ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ – ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

Ανάλυση Ποιότητας Δεδομένων: Μέθοδοι & Εφαρμογές

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Κ. Ντούσκας

Επιβλέπων: Κωνσταντίνος Δεμέστιχας, Επίκουρος Καθηγητής Γ.Π.Α.

Αθήνα, Οκτώβριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΕΠΙΣΤΗΜΟΝΙΚΟ – ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

Ανάλυση Ποιότητας Δεδομένων: Μέθοδοι & Εφαρμογές

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Κ. Ντούσκας

Επιβλέπων: Κωνσταντίνος Δεμέστιχας, Επίκουρος Καθηγητής Γ.Π.Α.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Οκτωβρίου 2024.

.....	.....	.....
Κωνσταντίνος Δεμέστιχας Επίκ. Καθηγητής Γεωπονικού Πανεπιστημίου Αθηνών	Ευγενία Αδαμοπούλου Ε.ΔΙ.Π. ΕΜΠ	Ευστάθιος Συκάς Ομότιμος Καθηγητής ΕΜΠ

Αθήνα, Οκτώβριος 2024

.....

Γεώργιος, Κ. Ντούσκας
Διπλωματούχος μεταπτυχιακού προγράμματος: «Τεχνοοικονομικά Συστήματα» της σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Ε.Μ.Π.

Περίληψη

Στη σημερινή εποχή, την εποχή της πληροφορίας, καθημερινά πραγματοποιείται ταχύτατη αποστολή, λήψη και επεξεργασία τεράστιου όγκου δεδομένων. Η επεξεργασία ορθών δεδομένων, ώστε να προκύπτουν αληθή συμπεράσματα και αποφάσεις, είναι μεγίστης σημασίας για άτομα, επιχειρήσεις και οργανισμούς. Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία προσπαθεί να καταγράψει, να εξετάσει και να αναλύσει το πρόβλημα της ποιότητας των δεδομένων, τις διαστάσεις, τα πρότυπα και τους κανόνες. Στη συνέχεια, γίνεται παρουσίαση σύγχρονων μεθόδων ανάλυσης ποιότητας δεδομένων από τη διεθνή βιβλιογραφία, εστιάζοντας τόσο στο ερευνητικό όσο και στο κανονιστικό πεδίο, εκτελώντας και μια συνοπτική σύγκριση αυτών. Τέλος, επιχειρείται προσπάθεια παρουσίασης των σύγχρονων λογισμικών που εστιάζουν στην ποιότητα των δεδομένων και μια συνοπτική σύγκριση αυτών.

Λέξεις-κλειδιά: Ανάλυση, Δεδομένα, Διαστάσεις, Μέθοδοι, Ποιότητα δεδομένων.

Abstract

In today's age, the age of information, huge amounts of data are sent, received and processed at high speed every day. The processing of correct data, so that true conclusions and decisions can be made, is of the utmost importance for individuals, businesses and organizations. Current master's thesis attempts to document, examine and analyze the problem of data quality, dimensions, standards and norms. Then, modern data quality analysis methods from the international literature are presented, focusing on both the research and regulatory fields, performing a brief comparison of them. Finally, an attempt is made to present modern software that focuses on data quality and a brief comparison of them.

Keywords: Analysis, Data, Data quality, Dimensions, Methods.

Περιεχόμενα

Κεφάλαιο 1^{ο} – Εισαγωγή στην ανάλυση ποιότητας δεδομένων

Ενότητα 1.1 – Εισαγωγή στα δεδομένα και την ποιότητα δεδομένων

Τα στοιχεία δεδομένων ή δεδομένα (data) αποτελούν «φορείς» πληροφορίας που συλλέγονται, δέχονται επεξεργασία και στη συνέχεια συνεισφέρουν στην διαμόρφωση των αποφάσεων ατόμων, εταιρειών, οργανισμών και κυβερνήσεων. Το μεγαλύτερο πλεονέκτημά τους είναι η δυνατότητα αποθήκευσης και ανάκτησης σε οποιοδήποτε χρόνο. Ιδιαίτερα στις μέρες μας, στη λεγόμενη «ψηφιακή εποχή», λόγω της έκρηξης της επιστήμης της πληροφορικής τα τελευταία 40 χρόνια και της δυνατότητας για ταχύτατη μεταφορά δεδομένων, πλήθος ιατρών, επιστημόνων, οικονομολόγων, μηχανικών και όχι μόνον, καθημερινά δέχονται και επεξεργάζονται μεγάλο όγκο δεδομένων (επεξεργασμένων ή ανεπεξέργαστων, αριθμητικών ή λεκτικών), που τους βοηθά στην εργασία τους.

Μέχρι τη δεκαετία του 1980, μαζικά κέντρα δεδομένων συγκέντρωναν δεδομένα, τα οποία αφορούσαν κατά κύριο λόγο την εξυπηρέτηση των επιχειρήσεων. Μέχρι το 2000, τα κέντρα δεδομένων επέκτειναν τη δυναμική τους στη συγκέντρωση δεδομένων, ώστε ολοένα και περισσότερα άτομα να έχουν δυνατότητα πρόσβασης σε προσωπικούς υπολογιστές και στο Διαδίκτυο (WWW/ World Wide Web) [1]. Κατά τη διάρκεια της πρώτης δεκαετίας του 2000 κι έπειτα με την ταυτόχρονη αύξηση της ταχύτητας του Διαδικτύου, τα κέντρα δεδομένων επέκτειναν τις δυνατότητές τους στην υποστήριξη της τεχνολογίας σύννεφου (cloud computing), με αποτέλεσμα τη ραγδαία αύξηση των συλλεγόμενων δεδομένων, τα οποία ήταν πλεόν διαθέσιμα σε όλους τους χρήστες.

Η έννοια «ποιότητα» (σύμφωνα με τον κανονισμό ISO 9000:2015) [2], είναι ο βαθμός στον οποίο ένα σύνολο εγγενών χαρακτηριστικών ενός αντικειμένου πληροί κάποιες υφιστάμενες απαιτήσεις. Ωστόσο, η βιβλιογραφία προσφέρει πρόσθετες ερμηνείες για την έννοια της ποιότητας δεδομένων, όπως για παράδειγμα, αν τα δεδομένα είναι κατάλληλα για τη χρήση που προορίζονται, και μάλιστα χωρίς κάποιου είδους ελάττωμα [3] ή ανταποκρίνονται στις ανάγκες και τις απαιτήσεις των χρηστών τους [4,5]. Στα παραπάνω πρέπει να ληφθεί υπόψιν ότι οι απαιτήσεις ποιότητας ενδέχεται

να επιβάλονται από πρότυπα, νομοθεσία, κανονισμούς, πολιτικές, ενδιαφερόμενα μέρη ή τη χρήση για την οποία προορίζονται [6].

Η ευρεία διαθεσιμότητα της σύγχρονης τεχνολογίας πληροφοριών, όπως οι smart συσκευές (κινητά τηλέφωνα, tablet και φορητές συσκευές), έχει ωθήσει τους ανθρώπους να παρακολουθούν τη φυσική τους κατάσταση, δραστηριότητα και άλλα δεδομένα υγείας ή διατροφικές συνήθειες ως χόμπι [7,8]. Οι εταιρείες έχουν μετατρέψει τα δεδομένα σε επιχειρηματικό μοντέλο (για παράδειγμα Google ή Meta) ή συσσωρεύουν δεδομένα για διαχείριση της γνώσης. Επιπλέον, συγκεκριμένοι επιστημονικοί κλάδοι, όπως η επιδημιολογία, αποκτούν δεδομένα για την έρευνα των συνθηκών υγείας και τα αίτιά τους [9]. Αυτά είναι μόνο μερικά παραδείγματα του πώς τα δεδομένα έχουν γίνει μέρος της καθημερινότητας. Ωστόσο, η πανταχού παρουσία των δεδομένων συμβαδίζει με την πανταχού παρουσία του ζητήματος ποιότητας αυτών. Απλά παραδείγματα από την καθημερινή ζωή είναι οι μη καταχωρημένες αλλαγές κατοικίας σε έναν κατάλογο επαφών, οι οποίες μπορεί να οδηγήσουν σε αδυναμία επικοινωνίας με ένα συγκεκριμένο άτομο ή μεροληπτικές στατιστικές αναλύσεις που λαμβάνουν υπόψη γεωγραφικές μεταβλητές, αμφισβητώντας τη χρησιμότητα του καταλόγου.

Η ποιότητα και η δυνατότητα εφαρμογής των δεδομένων θα πρέπει να εξετάζεται και αξιολογείται με προσοχή, καθώς μπορεί να επηρεάσουν σε σημαντικό βαθμό τη μετέπειτα επεξεργασία των δεδομένων, καθώς και τα αποτελέσματα ή τα συμπεράσματα που προκύπτουν από αυτά [10,11].

Η έρευνα και οι αναλύσεις υψηλής ποιότητας χρειάζονται αξιόπιστα δεδομένα. Ακόμα κι αν οι ποιοτικοί προβληματισμοί σχετικά με τα δεδομένα που συλλέγονται μπορεί να είναι τόσο παλιοί όσο και η ίδια η διαδικασία συλλογής, ουσιαστικά μόνο στη σύγχρονη βιβλιογραφία γίνεται εκτενής ανάλυση επί του ζητήματος [12,13]. Παρ 'όλα αυτά, η ποιότητα των δεδομένων είχε ήδη από το 1986 χαρακτηριστεί ως «ένα βασικό ζήτημα της εποχής μας» [14].

Ο πρωταρχικός στόχος του τομέα της ποιότητας δεδομένων είναι γενικά η διασφάλιση της ακεραιότητας των δεδομένων και επομένως, η διασφάλιση της χρηστικότητας και της χρησιμότητας αυτών. Ως εκ τούτου, οι ενδιαφερόμενοι για την ποιότητα των δεδομένων είναι οι παραγωγοί δεδομένων, οι χρήστες δεδομένων, οι αναλυτές και τα

άτομα που εξάγουν συμπεράσματα από «μεταφρασμένα» δεδομένα, όπως οι αποδέκτες των πληροφοριών που παρέχονται μέσω του διαδικτύου. Οι εκτιμήσεις για την ποιότητα των δεδομένων θα πρέπει να αφορούν πρωτίστως τα άτομα που συμμετέχουν στη συλλογή δεδομένων, τη δημιουργία δεδομένων ή τα άτομα που αναλύουν ή παρέχουν δεδομένα καθώς και τα άτομα με άμεση πρόσβαση σε αυτά, καθώς έχουν τα μέσα και τους μηχανισμούς να αντιμετωπίσουν ζητήματα ποιότητας δεδομένων. Ιδανικά, οι εκτιμήσεις ποιότητας δεδομένων προηγούνται και συνοδεύουν τη φάση συλλογής τους και μπορεί να συνεπάγονται, για παράδειγμα, κανόνες για τη διασφάλιση της δομής δεδομένων ή του εύρους τιμών. Ωστόσο, η διασφάλιση ποιότητας είναι μια συνεχής διαδικασία.

Προκειμένου να αποσαφηνιστούν τα παραπάνω, στόχοι της παρούσας εργασίας είναι:

- Η ανάλυση των τύπων των δεδομένων και του ζητήματος της ποιότητας δεδομένων.
- Η παρουσίση των διαστάσεων ποιότητας δεδομένων.
- Η παρουσίαση των Ευρωπαϊκών Οδηγιών καθώς και του Προτύπου ISO στην ποιότητα δεδομένων.
- Η ανασκόπηση των μεθόθων ανάλυσης και αξιολόγησης της ποιότητας δεδομένων στην υπάρχουσα βιβλιογραφία, συνοδευόμενη από μια συνοπτική σύγκριση.
- Η παρουσίαση, ανάλυση και σύγκριση των λογισμικών ποιότητας δεδομένων.

Ενότητα 1.2 – Κατηγοριοποίηση δεδομένων

Τα δεδομένα είναι αντικείμενα του πραγματικού κόσμου, με δυνατότητα αποθήκευσης, ανάκτησης και επεξεργασίας μέσω λογισμικών και μπορούν να μετακινούνται μέσω δικτύου [15]. Οι ερευνητές έχουν παρουσιάσει διαφορετικές ταξινομήσεις για τα δεδομένα σε διαφορετικές περιοχές έρευνας. Όμως, γενικά, τρεις τύποι δεδομένων περιγράφονται στο πεδίο της Ποιότητας Δεδομένων [16] βάσει της δομής τους. Ο Πίνακας 1.1 παρουσιάζει τύπους δεδομένων με βάση αυτή την ταξινόμηση.

Πίνακας 1.1 Τύποι δεδομένων βάσει της δομής τους.
(Πηγή:[17])

Είδη δεδομένων	Ορισμός	Παράδειγμα
Δομημένα δεδομένα (Structured data)	Γενίκευση ή συνάθροιση στοιχείων που περιγράφονται από στοιχειώδη χαρακτηριστικά που ορίζονται σε έναν τομέα.	Σχετικοί πίνακες Στατιστικά δεδομένα
Αδόμητα δεδομένα (Unstructured data)	Γενική ακολουθία συμβόλων, τυπικά κωδικοποιημένη σε φυσική γλώσσα.	Το σώμα ενός email Ερωτηματολόγιο με απαντήσεις ανοικτού τύπου
Ημι-δομημένα δεδομένα (Semi structure data)	Δεδομένα που έχουν δομή με κάποιο βαθμό ευελιξίας.	Γλώσσα σήμανσης (Markup language) π.χ. XML

Μια δεύτερη ταξινόμηση δεδομένων βασίζεται στην εξέταση των δεδομένων ως ένα «προϊόν». Αυτό το μοντέλο ταξινομεί τα δεδομένα σε τρεις τύπους. Ο Πίνακας 1.2 παρουσιάζει αυτή την ταξινόμηση.

Πίνακας 1.2 Τύποι δεδομένων βάσει της εξέτασης ως «προϊόν».
(Πηγή: [17])

Είδη δεδομένων	Ορισμός
Ακατέργαστα στοιχεία δεδομένων (Raw data items)	Μικρότερες ενότητες δεδομένων που χρησιμοποιούνται για τη δημιουργία πληροφοριών και στοιχείων δεδομένων.
Δεδομένα τύπου «εξαρτήματος» (Component data items)	Δεδομένα που κατασκευάζονται από ακατέργαστα στοιχεία δεδομένων και αποθηκεύονται προσωρινά μέχρι την κατασκευή του τελικού προϊόντος.
Προϊόντα πληροφορίας (Information products)	Δεδομένα, τα οποία είναι αποτέλεσμα της εκτέλεσης παραγωγικής δραστηριότητας σε δεδομένα.

Ένας άλλος τρόπος ταξινόμησης των δεδομένων βασίζεται στην αυστηρότητα ακριβούς μέτρησης και την επίτευξη της ποιότητας των δεδομένων, η οποία έχει δύο κατηγορίες: ειδικά στοιχειώδη δεδομένα και συγκεντρωτικά δεδομένα. Σε έναν οργανισμό, τα δεδομένα που διαχειρίζονται με λειτουργική διαδικασία και

αντιπροσωπεύουν ατομικά στοιχεία του πραγματικού κόσμου ονομάζονται στοιχειώδη δεδομένα (π.χ. φύλο, ηλικία), ενώ τα δεδομένα που προκύπτουν από στοιχειώδη δεδομένα για την εφαρμογή της συνάρτησης συνάθροισης ονομάζονται συγκεντρωτικά δεδομένα, όπως για παράδειγμα το μέσο εισόδημα που κατέβαλε ο φορολογούμενος σε μια συγκεκριμένη πόλη [15]. Τα δεδομένα μπορούν να ταξινομηθούν σε διαφορετικούς τύπους με βάση τη χρήση τους σε διάφορους τομείς (π.χ. δίκτυο ή ιστός/διαδίκτυο).

Ενότητα 1.3 – Κατηγοριοποίηση προβλημάτων ποιότητας δεδομένων

Το πρόβλημα της ποιότητας των δεδομένων γενικά μπορεί να διαχωριστεί σε δύο κατηγορίες: το πρόβλημα μιας πηγής και το πρόβλημα των πολλαπλών πηγών. Σύμφωνα με ορισμένες έρευνες, προσδιορίζονται τέσσερις κατηγορίες για την ποιότητα των δεδομένων, οι οποίες φαίνονται στον παρακάτω πίνακα (Πίνακας 1.3). Ως εκ τούτου, ο στόχος της ταξινόμησης του προβλήματος ποιότητας δεδομένων είναι ο εντοπισμός της ακριβούς εφαρμογής των δεδομένων για τις αντίστοιχες απαιτήσεις [18].

Πίνακας 1.3 Κατηγορίες προβλήματος ποιότητας δεδομένων.
(Πηγή: [17])

Πρόβλημα ποιότητας δεδομένων	**Κατηγορία**	**Ορισμός**
Πρόβλημα μιας πηγής (Single -source problem)	Επίπεδο Σχήματος	Έλλειψη περιορισμών ακεραιότητας - κακός σχεδιασμός Περιορισμοί μοναδικότητας Αναφορική ακεραιότητα
	Επίπεδο Στιγμιοτύπου	Σφάλματα εισαγωγής δεδομένων Ανορθογραφίες Διπλότυπα πλεονασμού Αντιφατικές αξίες
Πρόβλημα πολλαπλών πηγών (Multi-source problem)	Επίπεδο Σχήματος	Ετερογενή μοντέλα δεδομένων και σχεδιασμός σχημάτων Αντικρουόμενες καταχωρήσεις
	Επίπεδο Στιγμιοτύπου	Επικαλυπτόμενα αντιφατικά και ασυνεπή δεδομένα Ασυνεπής συγκέντρωση/συμψηφισμός Ασυνεπής συγχρονισμός

Ενότητα 1.4 – Διαστάσεις ποιότητας δεδομένων

Ο Πίνακας 1.4 απεικονίζει ορισμένες διαστάσεις ποιότητας δεδομένων και τον ορισμό τους από τη βιβλιογραφία. Από ερευνητικής πλευράς, παρουσιάζονται διαφορετικοί αριθμοί διαστάσεων για την Ποιότητα των Πληροφοριών και την Ποιότητα των Δεδομένων. Στην πραγματικότητα, η «Ποιότητα Δεδομένων», τα «Πληροφοριακά Συστήματα» και η «Λογιστική & Έλεγχος» είναι τρεις αρχικές κατηγορίες για τον προσδιορισμό των κατάλληλων διαστάσεων ποιότητας δεδομένων [19].

Στην κατηγορία της «Ποιότητας Δεδομένων», ο Wang [19] προσδιόρισε τέσσερις κατηγορίες που είναι οι: Ενδογένεια (Intrinsic), Προσβασιμότητα (Accessibility), Συνάφεια (Contextual) και Αντιπροσωπευτικότητα (Representational), καθώς και δεκαπέντε διαστάσεις για την Ποιότητα των Πληροφοριών (π.χ. αντικειμενικότητα, αξιοπιστία, φήμη, προστιθέμενη αξία). Άλλοι ερευνητές αναγνώρισαν επιπλέον διαστάσεις για την «Ποιότητα Δεδομένων», όπως επικύρωση δεδομένων, αξιοπιστία, ιχνηλασιμότητα, διαθεσιμότητα για ταυτοποίηση.

Στην κατηγορία των «Πληροφοριακών Συστημάτων», εντόπιστηκαν διάφοροι παράγοντες όπως η αξιοπιστία, η ακρίβεια, η συνάφεια, η χρηστικότητα και η ανεξαρτησία. Στην κατηγορία «Λογιστική & Έλεγχος», η ακρίβεια, η χρονική ισχύς των δεδομένων και η συνάφεια θεωρούνται τρεις διαστάσεις ποιότητας δεδομένων. Επιπρόσθετα, σε αυτή την κατηγορία ορισμένοι μελετητές αναφέρουν ότι τα συστήματα εσωτερικού ελέγχου χρειάζονται χαμηλότερο κόστος και υψηλότερη αξιοπιστία που αναφέρεται σε ορισμένες διαστάσεις όπως η ακρίβεια, η συχνότητα και το μέγεθος των δεδομένων [20].

Με βάση το πρότυπο ISO, όπως θα αναφερθεί και σε επόμενο κεφάλαιο, ποιότητα σημαίνει το σύνολο των χαρακτηριστικών μιας οντότητας που επηρεάζουν την ικανότητά της να ικανοποιεί προκαθορισμένες και υπονοούμενες ανάγκες [21]. Μια διάσταση ποιότητας δεδομένων στην πραγματικότητα, προσφέρει έναν τρόπο για τη μέτρηση και τη διαχείριση της ποιότητας των δεδομένων και των πληροφοριών [22]. Επομένως, το πρωταρχικό βήμα για την κατανόηση της διάστασης ποιότητας δεδομένων δρα επικουρικά και οδηγεί στη βελτίωση αυτής. Αναλυτές και προγραμματιστές χρησιμοποιούν τη διάσταση και την ταξινόμηση χωριστών δεδομένων, μέσω της χρήσης εργαλείων ποιότητας δεδομένων, για τη δημιουργία και

την επεξεργασία των πληροφοριών, προκειμένου να βελτιώσουν τις πληροφορίες και τη διαδικασία τους.

Πίνακας 1.4 Διαστάσεις ποιότητας δεδομένων.

(Πηγή: [17])

Διάσταση	**Ορισμός μέτρου διάστασης**
Επικαιρότητα ή Διαχρονική αξία δεδομένων (Timeliness)	Ο βαθμός στον οποίο η ηλικία των δεδομένων είναι κατάλληλη για την εκάστοτε εργασία. Η επικαιρότητα αναφέρεται μόνο στην καθυστέρηση μεταξύ της αλλαγής μιας κατάστασης του πραγματικού κόσμου και της επακόλουθης τροποποίησης της κατάστασης του συστήματος πληροφοριών. Η επικαιρότητα έχει δύο συνιστώσες: ηλικία και αστάθεια. Η ηλικία είναι ένα μέτρο της παλαιότητας των πληροφοριών. Η μεταβλητότητα είναι ένα μέτρο της αστάθειας της πληροφορίας της συχνότητας αλλαγής της τιμής για ένα χαρακτηριστικό μιας οντότητας.
Μοντερνισμός (Currency)	Ο μοντερνισμός είναι ο βαθμός στον οποίο ένα δεδομένο είναι ενημερωμένο. Μια τιμή αναφοράς είναι ενημερωμένη εάν είναι σωστή παρά τις πιθανές αποκλίσεις που προκαλούνται από αλλαγές που σχετίζονται με το χρόνο στη σωστή τιμή. Ο μοντερνισμός περιγράφει πότε οι πληροφορίες εισήχθησαν στις πηγές ή/και στην αποθήκη δεδομένων. Η μεταβλητότητα περιγράφει τη χρονική περίοδο για την οποία ισχύουν οι πληροφορίες.
Συνοχή (Consistency)	Ο βαθμός στον οποίο τα δεδομένα παρουσιάζονται με την ίδια μορφή και είναι συμβατικά με τα προηγούμενα δεδομένα. Αναφέρεται στην παραβίαση σημασιολογικών κανόνων που ορίζονται στο σύνολο των δεδομένων.
Ακρίβεια (Accuracy)	Τα δεδομένα είναι ακριβή, όταν οι τιμές δεδομένων που είναι αποθηκευμένες στη βάση δεδομένων αντιστοιχούν σε πραγματικές τιμές. Αφορά τον βαθμό στον οποίο τα δεδομένα είναι σωστά, αξιόπιστα και πιστοποιημένα. Η ακρίβεια είναι ένα μέτρο της εγγύτητας μιας τιμής δεδομένων, v, σε κάποια άλλη τιμή, v', που θεωρείται σωστή. Ένα μέτρο διόρθωσης των δεδομένων (το οποίο απαιτεί μια έγκυρη πηγή αναφοράς για να προσδιοριστεί).
Πληρότητα (Completeness)	Η ικανότητα ενός πληροφοριακού συστήματος να αναπαριστά κάθε ουσιαστική κατάσταση του πραγματικού κόσμου. Ο βαθμός στον οποίο τα δεδομένα είναι επαρκή για την εκάστοτε εργασία. Ο βαθμός στον οποίο ποιες τιμές υπάρχουν σε μια συλλογή δεδομένων. Ποσοστό των πληροφοριών του πραγματικού κόσμου που εισάγονται στις πηγές ή/και στην αποθήκη δεδομένων.
	Πληροφορίες που περιέχουν όλα τα απαιτούμενα μέρη των πληροφοριών μιας οντότητας. Αναλογία μεταξύ του αριθμού των μη μηδενικών τιμών σε μια πηγή και συνολικού μεγέθους των δεδομένων. Όλες οι τιμές που υποτίθεται ότι συλλέγονται σύμφωνα με μια θεωρία συλλογής.
Προσβασιμότητα (Accessibility)	Ο βαθμός στον οποίο οι πληροφορίες είναι διαθέσιμες ή ανακτώνται εύκολα και γρήγορα.
Διπλότυπα (Duplication)	Ένα μέτρο ανεπιθύμητης αντιγραφής που υπάρχει εντός ή μεταξύ συστημάτων για ένα συγκεκριμένο πεδίο, εγγραφή ή σύνολο δεδομένων.
Προδιαγραφές δεδομένων (Data specification)	Μέτρο ύπαρξης, πληρότητας, ποιότητας και τεκμηρίωσης προτύπων δεδομένων, μοντέλων δεδομένων, επιχειρηματικών κανόνων, μεταδεδομένων και δεδομένων αναφοράς.
Ποιότητας παρουσίασης (Presentation Quality)	Ένα μέτρο του τρόπου με τον οποίο παρουσιάζονται και συλλέγονται οι πληροφορίες. Η μορφή και η εμφάνιση βοηθούν στην κατάλληλη χρήση των πληροφοριών.
Συνεπής Εκπροσώπηση (Consistent Representation)	Ο βαθμός στον οποίο τα δεδομένα παρουσιάζονται στο ίδιο πλαίσιο/μορφή.
Φήμη (Reputation)	Ο βαθμός στον οποίο οι πληροφορίες τυγχάνουν μεγάλης εκτίμησης ως προς την πηγή ή το περιεχόμενο.
Ασφάλεια (Safety)	Είναι η ικανότητα της λειτουργίας να επιτυγχάνει αποδεκτά επίπεδα κινδύνου για τους ανθρώπους, τη διαδικασία, την ιδιοκτησία ή το περιβάλλον.
Επάρκεια δεδομένων (Appropriate amount of data)	Σε ποιο βαθμό ο όγκος των δεδομένων είναι κατάλληλος για την εκάστοτε εργασία.
Ασφάλεια (Security)	Ο βαθμός στον οποίο η πρόσβαση σε πληροφορίες περιορίζεται κατάλληλα για τη διατήρηση της ασφάλειάς τους.
Αξιοπιστία (Believability)	Ο βαθμός στον οποίο οι πληροφορίες θεωρούνται αληθείς και αξιόπιστες.
Κατανοησιμότητα (Understandability)	Ο βαθμός στον οποίο τα δεδομένα είναι ξεκάθαρα χωρίς ασάφεια και ευκόλως κατανοητά.
Αντικειμενικά (Αντικειμενικότητα) Objectively (Objectivity)	Ο βαθμός στον οποίο οι πληροφορίες είναι αμερόληπτες.

Σχετικότητα (Relevancy)	Ο βαθμός στον οποίο οι πληροφορίες είναι εφαρμόσιμες και χρήσιμες για την εκάστοτε εργασία.
Αποτελεσματικότητα (Effectiveness)	Είναι η ικανότητα της λειτουργίας να δίνει τη δυνατότητα στους χρήστες να επιτύχουν καθορισμένους στόχους με ακρίβεια και πληρότητα σε ένα συγκεκριμένο πλαίσιο χρήσης.
Μεταφρασιμότητα (Interpretability)	Ο βαθμός στον οποίο τα δεδομένα είναι δυνατόν να μεταφραστούν και τα σύμβολα, οι μονάδες και ο ορισμός είναι ξεκάθαροι.
Ευχρηστία (Ease of Manipulation)	Ο βαθμός στον οποίο τα δεδομένα είναι εύκολο να χρησιμοποιηθούν και να εφαρμοστούν σε διαφορετική ίδια μορφή.
Τελειότητα (Free-of-error)	Ο βαθμός στον οποίο τα δεδομένα είναι σωστά και αξιόπιστα.
Ευκολία χρήσης και δυνατότητα συντήρησης	Ένα μέτρο του βαθμού πρόσβασης και χρήσης των δεδομένων και του βαθμού στον οποίο τα δεδομένα μπορούν να ενημερωθούν, να διατηρηθούν και να διαχειριστούν.
Χρηστικότητα (Useability)	Στο βαθμό στον οποίο οι πληροφορίες είναι σαφείς και χρησιμοποιούνται εύκολα.
Αξιοπιστία (Reliability)	Ο βαθμός στον οποίο οι πληροφορίες είναι σωστές και αξιόπιστες. Είναι η ικανότητα της λειτουργίας να διατηρεί ένα συγκεκριμένο επίπεδο απόδοσης όταν χρησιμοποιείται σε καθορισμένες συνθήκες.
Ποσότητα δεδομένων (Amount of data)	Ο βαθμός στον οποίο η ποσότητα ή ο όγκος των διαθέσιμων δεδομένων είναι κατάλληλα.
Φρεσκάδα (Freshness)	Η «φρεσκάδα» αντιπροσωπεύει μια οικογένεια παραγόντων ποιότητας που ο καθένας αντιπροσωπεύει κάποια πτυχή «φρεσκάδας» την οποία διατηρεί και στις μετρήσεις του.
Προστιθέμενη αξία (Value added)	Ο βαθμός στον οποίο οι πληροφορίες είναι ωφέλιμες, παρέχει πλεονεκτήματα από τη χρήση τους.
Ικανότητα εκμάθησης (Learn ability)	Αφορά την ικανότητα της λειτουργίας να προσφέρει τη δυνατότητα στον χρήστη να την κατανοήσει.

Φθορά δεδομένων (Data Decay)	Ένα μέτρο του ρυθμού αρνητικής αλλαγής στα δεδομένα.
Περιεκτικότητα (Concise)	Ο βαθμός στον οποίο οι πληροφορίες αναπαρίστανται συμπαγώς, χωρίς να είναι ελλιπείς (δηλαδή συνοπτικές στην παρουσίαση, αλλά πλήρεις και επί της ουσίας).
Συνέπεια και συγχρονισμός (Consistency and Synchronization)	Ένα μέτρο της ισοδυναμίας των πληροφοριών που χρησιμοποιούνται σε διάφορες αποθήκες δεδομένων, εφαρμογές και συστήματα καθώς και τις διαδικασίες για την ισοδυναμία των δεδομένων.
Βασικές αρχές ακεραιότητας δεδομένων (Data integrity fundamentals)	Ένα μέτρο της ύπαρξης, της εγκυρότητας, της δομής, του περιεχομένου και άλλων βασικών χαρακτηριστικών των δεδομένων.
Πλοήγηση (Navigation)	Βαθμός στον οποίο τα δεδομένα εντοπίζονται εύκολα και συνδέονται εύκολα μεταξύ τους.
Χρησιμότητα (Useful)	Ο βαθμός στον οποίο οι πληροφορίες είναι εφαρμόσιμες και χρήσιμες για την εκάστοτε εργασία.
Αποδοτικότητα (Efficiency)	Ο βαθμός στον οποίο τα δεδομένα είναι σε θέση να καλύψουν γρήγορα τις ανάγκες πληροφοριών για την εκάστοτε εργασία.
Διαθεσιμότητα (Availability)	Ο βαθμός στον οποίο οι πληροφορίες είναι προσβάσιμες.
Κάλυψη δεδομένων (Data Coverage)	Ένα μέτρο της διαθεσιμότητας και της πληρότητας των δεδομένων σε σύγκριση με το πλήθος δεδομένων ή τον πληθυσμό ενδιαφέροντος.
Δυνατότητα Συναλλαγών (Transactability)	Ο βαθμός στον οποίο τα δεδομένα παράγουν την επιθυμητή επιχειρηματική συναλλαγή ή αποτέλεσμα.
Επικαιρότητα και Διαθεσιμότητα (Timeliness and Availability)	Ο βαθμός στον οποίο τα δεδομένα είναι επίκαιρα και διαθέσιμα για χρήση, όπως καθορίζεται από το χρονικό πλαίσιο στο οποίο αναφέρεται η μέτρηση.

Στο δεύτερο κεφάλαιο της εργασίας, παρουσιάζονται οι Ευρωπαϊκές Οδηγίες 2021 για το ζήτημα της ποιότητας δεδομένων, αναλύοντας πλήρως τις διαστάσεις που περιλαμβάνονται. Στο τρίτο κεφάλαιο θα παρουσιαστεί το πρότυπο Ποιότητας Δεδομένων ISO 25000. Στο τέταρτο κεφάλαιο, παρουσιάζονται σύγχρονες μεθόδοι ανάλυσης δεδομένων από τη διεθνή βιβλιογραφία, εστιάζοντας τόσο στο ερευνητικό όσο και στο κανονιστικό πεδίο, καθώς και συνοπτική σύγκριση αυτών. Στο πέμπτο κεφάλαιο, επιχειρείται προσπάθεια παρουσίασης των σύγχρονων λογισμικών που

εστιάζουν στην ποιότητα των δεδομένων (και όχι μόνον) και μια συνοπτική σύγκριση αυτών.

Κεφάλαιο 2^{ο} – Ευρωπαϊκές Οδηγίες 2021

Τον Αύγουστο του έτους 2021, εκδόθηκαν οι Ευρωπαϊκές Οδηγίες 2021 [23] που ορίζουν ένα συμβουλευτικό πλαίσιο για την επίτευξη της ποιότητας δεδομένων. Το εν λόγω πλαίσιο αφορά 4 διαστάσεις με τα αρχικά FAIR (Findability (Ευρεσιμότητα), Accessibility (Προσβασιμότητα), Interoperability (Διαλειτουργικότητα), Reusability (Επαναχρησιμοποίηση)) στους οποίους περιέχονται 12 σχετικοί δείκτες ποιότητας, όπως φαίνεται χαρακτηριστικά στον Πίνακα 2.1.

Πίνακας 2.1 Διάστασεις ποιότητας δεδομένων (FAIR) και αντίστοιχοι δείκτες.
(Πηγή: [23])

Ευρεσιμότητα (Findability)	**Προσβασιμότητα (Accessibility)**	**Διαλειτουργικότητα (Interoperability)**	**Επαναχρησιμοποίηση (Reusability)**
Πληρότητα (Completeness)	Προσβασιμότητα (Accessibility) / Διαθεσιμότητα (Availability)	Συμμόρφωση (Conformity/Compliance)	Επικαιρότητα ή Διαχρονική αξία δεδομένων (Timeliness)
Ευρεσιμότητα (Findability)		Μηχανική Αναγνωσιμότητα/ Επεξεργασιμότητα (Machine readability/processability)	Συνοχή (Consistency)
		Εξωστρέφεια (Openness)	Ακρίβεια (Accuracy)
			Συνάφεια (Relevance)
			Κατανοησιμότητα (Understandability)
			Αξιοπιστία (Credibility)

Ενότητα 2.1 – Διάσταση ευρεσιμότητας (Findability dimension)

Η εν λόγω διάσταση εισάγει δύο δείκτες όσον αφορά την ποιότητα των δεδομένων: την Πληρότητα (Completeness) και την Ευρεσιμότητα (Findability).

Πληρότητα (Completeness)

Τα μεταδεδομένα είναι περιγραφικά δεδομένα. Για παράδειγμα σε ένα ηχητικό κομμάτι οι πληροφορίες σχετικά με τον καλλιτέχνη και το άλμπουμ θεωρούνται μεταδεδομένα, καθώς αυτές οι πληροφορίες δεν αποτελούν μέρος του πραγματικό αρχείου. Είναι, ωστόσο, πολύ σημαντικές πληροφορίες όταν προσπαθεί ένας χρήστης να βρει το αρχείο μεταξύ πολλών άλλων. Ομοίως, εάν από ένα έγγραφο κειμένου έλειπε ο τίτλος του θα ήταν πολύ δύσκολο για τους χρήστες να ανακαλύψουν το έγγραφο. Τα πλήρη

και ενημερωμένα μεταδεδομένα είναι, επομένως, ζωτικής σημασίας για την εύρεση και χρήση των δεδομένων.

Επιπλέον, τα μεταδεδομένα μπορούν να βοηθήσουν τους χρήστες να προσδιορίσουν, εάν οι πληροφορίες που ανακτήθηκαν αντιστοιχούν στην ανάγκη τους. Μια βιβλιοθήκη βιβλίων θα είχε ελάχιστη χρησιμότητα αν από τα βιβλία έλειπαν οι βασικές πληροφορίες μεταδεδομένων τους: για παράδειγμα συγγραφέας, τίτλος και ISBN. Το ίδιο ισχύει σε δεδομένα που δημοσιεύονται στο διαδίκτυο. Συχνά, όταν κάποιος δημοσιεύει τα δεδομένα του σε έναν κατάλογο, ορισμένα πεδία μεταδεδομένων ορίζονται ως υποχρεωτικά, πράγμα που σημαίνει ότι πρέπει να συμπληρωθούν για να μπορέσουν να δημοσιευτούν τα εν λόγω δεδομένα. Ωστόσο, συνιστάται τα πεδία μεταδεδομένων που δεν έχουν οριστεί ως υποχρεωτικά να συμπληρωθούν εξίσου. Για τον εκδότη των δεδομένων δεν χρειάζεται συνήθως μεγάλη προσπάθεια για τη συμπλήρωση αυτών των πεδίων, ενώ και για τους χρήστες δεδομένων τα πλήρη μεταδεδομένα μπορούν να είναι πολύ ωφέλιμα. Όσο περισσότερες πληροφορίες υπάρχουν σχετικά με τα δεδομένα, τόσο πιο εύκολο είναι για τους χρήστες να τα ανακαλύψουν και να αποκτήσουν μια πρώτη εικόνα τους, που με τη σειρά της αυξάνει τις πιθανότητες να τα ξαναχρησιμοποιήσουν.

Οι ακόλουθες πληροφορίες μεταδεδομένων θα πρέπει να παρέχονται προκειμένου να αυξηθεί διάσταση Ευρεσιμότητας των δεδομένων:

- Τίτλος
- Περιγραφή
- Λέξεις-κλειδιά
- Κατηγορίες
- Χρονικές πληροφορίες
- Χωρικές πληροφορίες.

Κατά τη συμπλήρωση αυτών των πληροφοριών μεταδεδομένων, οι εκδότες δεδομένων πρέπει να διασφαλίζουν ότι οι πληροφορίες που δίνονται είναι όσο το δυνατόν πιο ακριβείς και χρήσιμες, έτσι ώστε να είναι ξεκάθρο στον πιθανό χρήστη στο τι αφορούν τα εν λόγω δεδομένα.

Ευρεσιμότητα (Findability)

Σε αρκετές περιπτώσεις, τα δεδομένα δεν είναι πλήρη. Ωστόσο, μια τιμή που λείπει δεν αποτελεί λόγο, ώστε να μην γίνει δημοσίευση των υπόψη στοιχείων. Προκειμένου να αποφευχθεί η σύγχυση, ο πάροχος δεδομένων θα πρέπει να επισημαίνει σαφώς τις τιμές που λείπουν, ως μηδενικές τιμές. Οι χρήστες που δεν είναι εξοικειωμένοι με τα δεδομένα μπορούν με αυτό τον τρόπο, να αναγνωρίσουν ότι τα δεδομένα δεν ξεχάστηκαν απλώς, επειδή η τιμή μηδέν χρησιμοποιείται ως ειδικός δείκτης που υποδεικνύει ότι η τιμή δεν υπάρχει. Με άλλα λόγια, μια μηδενική τιμή είναι μια οπτική αναπαράσταση μιας τιμής που λείπει.

Υπάρχουν διάφοροι τρόποι για να δηλωθεί μια μηδενική τιμή, για παράδειγμα επισημαίνοντας την τιμή που λείπει με “NULL” ή “NA”. Ωστόσο, αν μέσα στα δεδομένα υπάρχει ένα υψηλό ποσοστό μηδενικών τιμών εντός μιας γραμμής ή στήλης, θα πρέπει να εξετάζεται το ενδεχόμενο διαγραφής της αντίστοιχης στήλης ή σειράς, καθώς πιθανότατα δεν αποφέρει προστιθέμενη αξία στους πιθανούς χρήστες των δεδομένων.

Ενότητα 2.2 – Διάσταση προσβασιμότητας (Accessibility dimension)

Η εν λόγω διάσταση εισάγει δύο συσχετιζόμενους δείκτες όσον αφορά την ποιότητα των δεδομένων: την Προσβασιμότητα (Accessibility) και τη Διαθεσιμότητα (Availability).

Προσβασιμότητα (Accessibility)

Μία από τις βασικές αρχές των ανοιχτών δεδομένων είναι η προσβασιμότητά τους: τα δεδομένα πρέπει να είναι προσβάσιμα και να διατίθεται στο ευρύτερο δυνατό φάσμα χρηστών για να αποφευχθεί ο περιορισμός της πιθανής επαναχρησιμοποίησής τους. Για να επιτρέπεται η εύκολη χρήση και η περαιτέρω επεξεργασία, δεν θα πρέπει να υπάρχουν περιορισμοί πρόσβασης, ανεξάρτητα από το αν απαιτούν χειροκίνητη παρέμβαση (π.χ. εγγραφή) ή αυτόματη παράκαμψη (π.χ. παροχή διαπιστευτηρίων). Αυτό ισχύει και για τα αρχεία τα ίδια, όπως για παράδειγμα τα κρυπτογραφημένα αρχεία. Οποιοσδήποτε περιορισμός πρόσβασης περιορίζει τον αριθμό των πιθανών χρηστών δεδομένων και επομένως, εάν είναι δυνατόν, θα πρέπει να αποφεύγεται, εκτός αν πρέπει να ακολουθηθούν συγκεκριμένοι κανόνες ή νόμοι.

Διαθεσιμότητα (Availability)

Τα δεδομένα μπορούν να επαναχρησιμοποιηθούν από άλλους χρήστες μόνο εάν είναι προσβάσιμα. Συνήθως, το κύριο σημείο πρόσβασης είναι μια διεύθυνση URL λήψης, η οποία πρέπει να οριστεί στα μεταδεδομένα και να είναι προσβάσιμη, για παράδειγμα μέσω ενός προγράμματος περιήγησης. Επομένως, ο εκδότης δεδομένων πρέπει να διασφαλίσει ότι, όταν ένας χρήστης κάνει “κλικ” στη διεύθυνση URL λήψης που παρέχεται, αυτή η διεύθυνση URL λειτουργεί σωστά και ο χρήστης απευθείας αποθηκεύει τα δεδομένα.

Ενότητα 2.3 – Διάσταση διαλειτουργικότητας (Interoperability dimension)

Η εν λόγω διάσταση αφορά τρεις δείκτες σχετικά με την ποιότητα των δεδομένων: τη Συμμόρφωση (Conformity/Compliance), τη Μηχανική Αναγνωσιμότητα/ Επεξεργασιμότητα (Machine readability/processability) και την Εξωστρέφεια (Openness).

Συμμόρφωση (Conformity/Compliance)

Τα δεδομένα (και τα μεταδεδομένα) συχνά περιέχουν ως πληροφορία ημερομηνίες και ώρες. Ανάλογα με την περιοχή/χώρα, παρουσιάζονται διαφορετικοί τρόποι αναφοράς ημερομηνιών, οι οποίοι μπορεί να οδηγήσουν σε σύγχυση. Το ακόλουθο παράδειγμα τονίζει το πρόβλημα με τις διφορούμενες μορφές ημερομηνίας: η 01/03/2021 θα μπορούσε σημαίνει είτε 1 Μαρτίου 2021 είτε 3 Ιανουαρίου 2021, ανάλογα με τη χώρα αναφοράς. Επομένως, η ημερομηνία και η ώρα θα πρέπει πάντα να κωδικοποιούνται ως ISO 8601 (EEEE-MM-HH & ωω:λλ:σσ). Θα πρέπει, επίσης, να αναφέρεται η ζώνη ώρας που χρησιμοποιείται. Η ζώνη ώρας πάντα προέρχεται από τη Συντονισμένη Παγκόσμια Ώρα (UTC).

Επίσης, τα δεδομένα περιέχουν συχνά αριθμούς. Για παράδειγμα, το κόμμα χρησιμοποιείται συχνά για τον διαχωρισμό ακέραιων αριθμών από δεκαδικούς. Αυτό μπορεί να προκαλέσει προβλήματα, όπως σε ένα αρχείο CSV όταν το διαχωριστικό μεταξύ των τιμών έχει οριστεί ως ένα κόμμα. Για να αποφευχθεί η ακούσια ερμηνεία ενός κόμματος που χωρίζει έναν αριθμό από δεκαδικό, θα πρέπει να χρησιμοποιηθεί μια τελεία. Σε μεγάλους αριθμούς, μερικές φορές χρησιμοποιείται διαχωριστικό χιλίων, για παράδειγμα μια τελεία ή κενό διάστημα. Αυτό μπορεί να οδηγήσει σε παρερμηνεία (ειδικά όταν τα δεδομένα υποβάλλονται σε αυτόματη επεξεργασία) και μπορεί να σημαίνει ότι ο χρήστης πρέπει να καθαρίσει τα δεδομένα πριν μπορέσουν να ξαναχρησιμοποιηθούν. Ως εκ τούτου, δεν πρέπει να χρησιμοποιούνται διαχωριστές για χιλιάδες.

Για να εξασφαλιστεί ότι οι χαρακτήρες παρουσιάζονται σωστά και να διασφαλιστεί η μεγαλύτερη δυνατή συμβατότητα με εφαρμογές που επεξεργάζονται δεδομένα, θα πρέπει πάντα να χρησιμοποιείται μια τυποποιημένη κωδικοποίηση χαρακτήρων. Συνήθως, το UTF-8 είναι η κωδικοποίηση της επιλογής στον Ιστό. Το UTF-8 είναι μια

κωδικοποίηση χαρακτήρων για το Unicode, ένα διεθνές πρότυπο για την αναπαράσταση όλων των χαρακτήρων με νόημα. Με αυτό, όλοι οι χαρακτήρες, ανεξαρτήτως γλώσσας, εμφανίζονται σωστά. Για να διασφαλιστεί ότι τα δεδομένα μπορούν να συνδυαστούν και να επαναχρησιμοποιηθούν με άλλα δεδομένα από διεθνείς πηγές και για να αποφευχθούν προβλήματα κατά την μηχανική επεξεργασία, είναι χρήσιμο να χρησιμοποιείται εξαρχής ένα διεθνώς αναγνωρισμένο και ευρέως χρησιμοποιούμενο σύνολο χαρακτήρων.

Ωστόσο, γενικά θα πρέπει να αποφεύγεται η χρήση ειδικών χαρακτήρων στα δεδομένα, ακόμα κι αν αποτελούν μέρος του UTF-8. Με αυτόν τον τρόπο, ενθαρρύνεται η προς τα πίσω συμβατότητα με παλαιότερα συστήματα.

Μηχανική αναγνωσιμότητα/επεξεργασιμότητα (Machine readability/ processability)

Ο εν λόγω δείκτης παρουσιάζει τρόπους, ώστε ο χρήστης να εισάγει τα δεδομένα με μορφή, ώστε να είναι σωστή η ανάγνωση των δεδομένων και εύκολη η επεξεργασία από το εκάστοτε σύστημα. Συγκεκριμένα:

1. Χρηση ενός ερωτηματικού ως οριοθέτη.
2. Χρήση ενός αρχείου ανά πίνακα.
3. Αποφυγή χρήσης κενού διαστήματος και πρόσθετων πληροφοριών.
4. Εισαγωγή κεφαλίδων στηλών.
5. Οι σειρές θα πρέπει να έχουν τον ίδιο αριθμό στηλών.
6. Αναφορά των μονάδων μέτρησης με χρηστικό τρόπο που να διευκολύνει την επεξεργασία των δεδομένων.

Εξωστρέφεια (Openness)

Η “εξωστρέφεια” των δεδομένων είναι ζωτικής σημασίας για την έννοια των Ανοιχτών Δεδομένων (Open Data). Τα δεδομένα θεωρούνται ανοιχτά, εάν οι πόροι είναι διαθέσιμοι σε μη αποκλειστική μορφή και μπορούν να χρησιμοποιηθούν με ανοιχτή άδεια.

Στις οδηγίες παρουσιάζεται ένα μοντέλο πέντε βημάτων – αστέρων, το οποίο μπορεί διαδοχικά να οδηγήσει σε ανοιχτά δεδομένα, καθώς και στη μέτρηση της εξωστρέφειάς τους.

Η "εξωστρέφεια" έχει ιδιαίτερη σημασία κατά τη δημοσίευση των δεδομένων. Επηρεάζει άμεσα την ικανότητα των χρηστών να επαναχρησιμοποιούν και να επεξεργάζονται δεδομένα και συνεπώς επηρεάζει την αξία των δεδομένων. Ιδιαίτερη σημασία δίνεται στην "εξωστρέφεια" όσον αφορά τις μορφές αρχείων. Το μοντέλο πέντε αστέρων του Tim Berners-Lee [24], το οποίο αναπτύχθηκε το 2001, είναι μια προσπάθεια να διαμορφωθεί μια κλίμακα για τη μέτρηση της "εξωστρέφειας" των δεδομένων. Οι τάξεις είναι κλιμακωτές, δηλαδή για να τηρηθεί μια συγκεκριμένη κατάταξη πρέπει να πληρούνται και τα κριτήρια των προηγούμενων βαθμών. Ανεξάρτητα από την πραγματική ποιότητα των δεδομένων, το πρώτο αστέρι απονέμεται για χρήση ανοιχτής άδειας. Εάν η χρήση δεδομένων περιορίζεται από μια ιδιόκτητη άδεια η ποιότητα της εξωστρέφειας καθίσταται άνευ νοήματος. Για να επιτευχθεί το δεύτερο αστέρι, η επιλεγμένη μορφή αρχείου πρέπει να είναι ημι-δομημένη. Ένας πίνακας που είναι αποθηκευμένος ως CSV, επεξεργάζεται από έναν χρήστη πιο εύκολα από μια εικόνα στην οποία απεικονίζεται ένας πίνακας. Για βαθμολόγηση με τρία αστέρια, απαιτείται η χρήση μη αποκλειστικών μορφών. Η χρήση URIs (Uniform Resource Identifiers) ως αναγνωριστικών για πόρους απαιτείται για βαθμολόγηση με 4 αστέρια. Το αποφασιστικό χαρακτηριστικό για την επίτευξη των πλήρων πέντε αστεριών είναι η σύνδεση δεδομένων μεταξύ τους για την παροχή του περιεχομένου. Μια απεικόνιση αυτής της ιεραρχίας φαίνεται στην Εικόνα 2.1.

Εικόνα 2.1 Μοντέλο πέντε αστέρων "εξωστρέφειας" δεδομένων του Tim Berners-Lee.
(Πηγή: [23])

Όπως αναφέρθηκε παραπάνω, το πρώτο αστέρι απονέμεται για χρήση ανοιχτής άδειας. Για να επιτευχθεί βαθμολογία δύο αστέρων, τα δεδομένα πρέπει να είναι δομημένα. Η Εικόνα 2.1 παρέχει μια επισκόπηση των κοινών μορφών και υποδεικνύει εάν είναι αναγνώσιμες από μηχανή ή όχι. Με βάση αυτό, οι προτεινόμενες μορφές για εκδότες δεδομένων είναι RDF, XML, JSON και CSV.

Η χρήση μιας μορφής αναγνώσιμης από μηχανή είναι το κλειδί για την επίτευξη υψηλού επιπέδου "εξωστρέφειας". Ωστόσο, ορισμένες μορφές, όπως το XLS, είναι ιδιόκτητες, πράγμα που σημαίνει ότι απαιτείται ένα συγκεκριμένο λογισμικό - σε αυτήν την περίπτωση το Microsoft Excel - για την πλήρη επεξεργασία του αρχείου. Συχνά, αυτού του είδους το λογισμικό δεν είναι ελεύθερα διαθέσιμο. Καθώς η προσβασιμότητα για όλους αποτελεί βασική αρχή των ανοιχτών δεδομένων, οι ιδιόκτητες μορφές αρχείων δεν είναι η σωστή επιλογή. Ως εκ τούτου, για βαθμολόγηση τριών αστέρων, πρέπει να χρησιμοποιείται μια μη ιδιόκτητη μορφή αρχείου, όπως το ODS ή το CSV.

Τα δεδομένα τριών αστέρων είναι εύκολα επεξεργάσιμα, αλλά απομονωμένα και δύσκολο να τα αναφέρουν άλλοι. Προκειμένου να επιτευχθεί βαθμολογία τεσσάρων αστέρων, URIs πρέπει να χρησιμοποιούνται για να υποδηλώνουν πόρους ή έννοιες. Φυσικά, το ίδιο το αρχείο θα πρέπει επίσης να αναφέρεται από ένα URI. Η σύσταση εστιάζει στη χρήση URI στα ίδια τα δεδομένα.

Τα γραφήματα RDF αποτελούνται από τριάδες, που αποτελούνται από υποκείμενο, κατηγορούμενο και αντικείμενο. Τα υποκείμενα και τα κατηγορούμενα πρέπει πάντα να είναι πόροι, ενώ τα αντικείμενα μπορεί να είναι είτε πόροι είτε απλές τιμές δεδομένων.

Προκειμένου να αντικατασταθούν τα αναγνωριστικά με URI, θα μπορούσαν να εξεταστούν τα υπαρχόντα ελεγχόμενα λεξιλόγια και οι βάσεις γνώσεων, για την περίπτωση που οι έννοιες έχουν ήδη ευρέως υιοθετήσει URIs. Εάν όχι, η αρχή που δημοσιεύει τα δεδομένα μπορεί να δημοσιεύσει τη δική της οντολογία/ορολογία, προκειμένου να ορίσει έννοιες που δεν έχουν καθοριστεί αλλού.

Το κύριο πλεονέκτημα της χρήσης URIs για τον προσδιορισμό πραγμάτων, είναι ότι μετατρέπει τις πληροφορίες σε αναφορικές. Δεδομένου, ότι ο Ιστός βασίζεται κυρίως στο πρωτόκολλο HTTP, τα URIs δεν είναι απλώς μοναδικά αναγνωριστικά, αλλά και άμεσα επιλύσιμα, υποδεικνύοντας έτσι τον πόρο. Το επόμενο βήμα είναι η πραγματική

σύνδεση των πακέτων πληροφορίας μεταξύ τους, προκειμένου να δημιουργηθούν συνδεδεμένα δεδομένα. Ένα σημασιολογικό γράφημα, γνωστό και ως γράφημα γνώσης, μπορεί να διαμορφωθεί μόνο χρησιμοποιώντας μορφή RDF. Ένα γράφημα που έχει κατασκευαστεί με αυτόν τον τρόπο μπορεί να διασχιστεί με επίλυση, δηλαδή απο-αναφορά, των HTTP URIs. Αυτό σημαίνει ότι μπορούν να συναχθούν δεδομένα και να ανακαλυφθούν περισσότερες σχέσεις. Τα δεδομένα εμπλουτίζονται με την προσθήκη αναφορών URIs σε άλλες πηγές. Μπορούν να δημιουργηθούν σύνδεσμοι, για παράδειγμα, με τα ελεγχόμενα λεξιλόγια που δημοσιεύονται από την Υπηρεσία Εκδόσεων ή την DBpedia. Η χρήση RDF και η σύνδεση δεδομένων απαιτούνται για την επίτευξη της λήψης βαθμολογίας πέντε αστέρων.

Ο Πίνακας 2.2 δείχνει μια λίστα των μορφών που χρησιμοποιούνται συνήθως, μαζί με πληροφορίες σχετικά με το εάν είναι μηχανικά αναγνώσιμες. Η δεξιά στήλη υποδεικνύει τον αριθμό των αστεριών που μπορούν να ληφθούν κατά τη χρήση αυτής της μορφής για δημοσίευση δεδομένων. Οι μορφές επιλέχθηκαν με βάση την ανάλυση που πραγματοποιήθηκε στη φάση δημιουργίας προφίλ δεδομένων. Στην ιδανική περίπτωση, θα πρέπει να χρησιμοποιούνται οι μορφές που επισημαίνονται με πράσινο χρώμα. Εάν αυτό δεν είναι δυνατό, θα πρέπει να χρησιμοποιηθούν μορφές από την κίτρινη ενότητα. Θα πρέπει να αποφεύγεται η χρήση μορφών που επισημαίνονται με κόκκινο χρώμα, καθώς με αυτές μπορεί να επιτευχθεί βαθμολογία μόνο ενός αστεριού.

Πίνακας 2.2 Τύπος αρχείων και επίπεδο "εξωστρέφειας" δεδομένων.
(Πηγή: [23])

Format	Non-proprietary	Machine readable	Achievable stars
RDF	Yes	Yes	✪ ✪ ✪ ✪ ✪
XML	Yes	Yes	✪ ✪ ✪ ✪
JSON	Yes	Yes	✪ ✪ ✪ ✪
CSV	Yes	Yes	✪ ✪ ✪ ✪
ODS	Yes	Predominantly	✪ ✪ ✪ ✪
XLSX	Yes	Predominantly	✪ ✪ ✪ ✪
XLS	No	Predominantly	✪ ✪
TXT	Yes	Predominantly	✪
HTML	Yes	Predominantly	✪
PDF	Yes	No	✪
DOCX	Yes	No	✪
ODT	Yes	No	✪
PNG	Yes	No	✪
GIF	No	No	✪
JPG/JPEG	No	No	✪
TIFF	No	No	✪
DOC	No	No	✪

Ενότητα 2.4 – Διάσταση επαναχρησιμοποίησης (Reusability dimension)

Η εν λόγω διάσταση αφορά έξι δείκτες όσον αφορά την ποιότητα των δεδομένων:

- Επικαιρότητα ή Διαχρονική αξία δεδομένων (Timeliness)
- Συνοχή (Consistency)
- Ακρίβεια (Accuracy)
- Συνάφεια (Relevance)
- Κατανοησιμότητα (Understandability)
- Αξιοπιστία (Credibility)

Επικαιρότητα ή Διαχρονική αξία Δεδομένων (Timeliness)

Τα μεταδεδομένα και τα δεδομένα είναι επίκαιρα, εάν είναι ενημερωμένα και αντιπροσωπεύουν την πραγματική και τρέχουσα κατάσταση. Αυτό σημαίνει ότι μόλις μια αλλαγή εμφανίζεται στον πραγματικό κόσμο, τα δεδομένα και τα μεταδεδομένα πρέπει να τροποποιηθούν επίσης. Ωστόσο, είναι δύσκολο να κατανοηθεί αυτόματα από το περιεχόμενο, αν τα δεδομένα είναι ιστορικά αληθή ή είναι αληθή σε πραγματικό χρόνο. Ως εκ τούτου, δεν έχουν καθοριστεί απαιτήσεις του συγκεκριμένου δείκτη με αυτοματοποιημένο τρόπο.

Συνοχή (Consistency)

Τα κοινοτικά πρότυπα είναι ένα ισχυρό εργαλείο για τη διασφάλιση της συμμόρφωσης μεταξύ αρχείων και μορφών ενός κοινού τομέα (domain). Η χρήση κοινοτικών προτύπων διευκολύνει την επαναχρησιμοποίηση δεδομένων, καθώς όλα τα δεδομένα που ακολουθούν το ίδιο πρότυπο “επικοινωνούν” - για παράδειγμα είναι οργανωμένα με τυποποιημένο τρόπο, η τεκμηρίωση ακολουθεί ένα κοινό πρότυπο ή χρησιμοποιείται ένα κοινό λεξιλόγιο. Υπάρχουν πολλά διαφορετικά κοινοτικά πρότυπα, όπως πρότυπα για συγκεκριμένους τομείς, π.χ. το κλίμα και οι προβλέψεις. Υπάρχουν όμως και πρότυπα που δεν αφορούν συγκεκριμένους τομείς, όπως το DCAT-AP, ένα πρότυπο για την αποθήκευση μεταδεδομένων καταλόγου δεδομένων.

Ανάλογα με την περίπτωση χρήσης, θα πρέπει να υπάρχει διασφάλιση που βοηθά στον έλεγχο των αρχείων, σύμφωνα με ένα τέτοιο πρότυπο. Η διασφάλιση της συμμόρφωσης των αρχείων με τα κοινοτικά πρότυπα βοηθά σημαντικά την επαναχρησιμοποίηση και διευκολύνει την περαιτέρω επεξεργασία.

Επιπρόσθετα, κάθε κομμάτι δεδομένων πρέπει να είναι μοναδικό. Τα διπλά δεδομένα δεν έχουν πρόσθετη αξία. Αντίθετα, μειώνουν την ποιότητα των δεδομένων, καθώς μπορεί να προκαλέσουν σφάλματα κατά την περαιτέρω επεξεργασία. Για παράδειγμα, ένας χρήστης δεδομένων που εκτελεί αναλύσεις σε δεδομένα με ορισμένες διπλότυπες εγγραφές, πιθανότατα θα λάβει λάθος αποτελέσματα.

Ακρίβεια (Accuracy)

Η ακρίβεια μπορεί να μετρηθεί σε πολλές διαστάσεις. Τι σημαίνει συγκεκριμένα ακρίβεια, πώς μετριέται και ποιο αποτέλεσμα θεωρείται αποδεκτό εξαρτώνται πάντα από τη συγκεκριμένη περίπτωση χρήσης. Για παράδειγμα, σε αρχεία .csv, κάθε κελί μιας στήλης θα μπορούσε να ελεγχθεί για ακρίβεια σε σχέση με μια μορφή κωδικοποίησης, για παράδειγμα βάσει ISO 8601:1988 [25], οι εγγραφές ημερομηνίας. Η αναλογία μεταξύ ακριβών και ανακριβών κελιών θα μπορούσε στη συνέχεια να δώσει στους χρήστες μια πρώτη εντύπωση της ακρίβειας των δεδομένων και πόσο δύσκολη μπορεί να είναι η επεξεργασία τους στη συνέχεια. Η υψηλότερη ακρίβεια είναι συνήθως ένας δείκτης δεδομένων υψηλότερης ποιότητας.

Κατά τη δημοσίευση δεδομένων, καλό είναι να παρέχονται και πληροφορίες για το μέγεθος byte των διανομών. Αυτές οι πληροφορίες βοηθούν τους χρήστες και τις αυτοματοποιημένες διαδικασίες να προβλέψουν τι να περιμένουν πριν από τη λήψη του πραγματικού αρχείου. Επίσης, αυτές οι πληροφορίες επιτρέπουν το φιλτράρισμα κατά μέγεθος.

Συνάφεια (Relevance)

Ανάλογα με τα προς δημοσίευση δεδομένα, η έννοια του όρου "κατάλληλο" μπορεί να διαφέρει πολύ. Είναι σημαντικό να δημοσιεύονται όλα τα σχετικά δεδομένα, αλλά θα πρέπει να δίνεται προσοχή να μην δημοσιεύονται τυφλά όλα τα διαθέσιμα δεδομένα χωρίς να λαμβάνεται υπόψη η χρησιμότητά τους. Από την άλλη πλευρά, οι εκδότες δεδομένων πρέπει να βεβαιωθούν ότι δημοσιεύεται επαρκής ποσότητα δεδομένων, ώστε να υπάρχει αρκετή πληροφορία και οι χρήστες να μπορούν να αντλούν αξία από αυτό. Ωστόσο, δεν υπάρχει σαφής ένδειξη για το ποια είναι η κατάλληλη ποσότητα δεδομένων, καθώς αυτό εξαρτάται σε μεγάλο βαθμό από τον σκοπό που έχει κατά νου ένας χρήστης.

Η ζήτηση μεγάλων ποσοτήτων δεδομένων μπορεί εύκολα να δημιουργήσει υψηλά φορτία στον διακομιστή. Σε ορισμένες περιπτώσεις, δεν απαιτούνται όλα τα δεδομένα ή τουλάχιστον όχι όλα ταυτόχρονα. Προκειμένου να μειωθεί αυτό το φορτίο και να αυξηθούν οι χρόνοι απόκρισης, θα πρέπει να χρησιμοποιείται σελιδοποίηση κατά περίπτωση. Αυτό σημαίνει ότι εμφανίζονται τμήματα δεδομένων αντί για το σύνολο των δεδομένων. Ο πελάτης μπορεί να δηλώσει στο αίτημα ποιο κομμάτι θα ανακτήσει, καθώς και το μέγεθός του. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας τις παραμέτρους που φαίνονται στον παρακάτω πίνακα.

Πίνακας 2.3 Παράμετροι βελτιστοποίησης φορτίου δεδομένων στον διακομιστή. (Πηγή [23])

Παράμετρος	Συμπεριφορά
Offset	Καθορίζει τον πόρο από τον οποίο θα ξεκινήσει η μέτρηση.
Όριο (Limit)	Καθορίζει πόσοι πόροι θα ανακτηθούν.

Κατανοησιμότητα (Understandability)

Τα δεδομένα και τα μεταδεδομένα είναι κατανοητά, εάν είναι σαφή και κατανοητά στον χρήστη. Μετά τη μελέτη των δεδομένων και των μεταδεδομένων, δεν πρέπει να υπάρχουν ασάφειες στον χρήστη.

Αυτός ο δείκτης εξαρτάται σε μεγάλο βαθμό από την αντίληψη του χρήστη και τις ειδικές γνώσεις του στον συγκεκριμένο τομέα. Ο βαθμός κατανοησιμότητας μπορεί να αυξηθεί, εάν παρέχονται συγκεκριμένες πληροφορίες σχετικά με τα συμφραζόμενα,

όπως περιγραφή των δεδομένων, τίτλος και λέξεις-κλειδιά. Ωστόσο, τελικά εξαρτάται από τον χρήστη, εάν τα δεδομένα είναι πραγματικά κατανοητά ή όχι.

Επίσης, τα APIs (Application Programming Interface) θα πρέπει να προσδιορίζονται όσο το δυνατόν λεπτομερέστερα. Αυτά περιλαμβάνουν διαθέσιμες διαδρομές, επιστρεφόμενες μορφές και κωδικούς κατάστασης. Εάν ένα API επιτρέπει τη μεταφόρτωση αρχείων, θα πρέπει επίσης να αναφέρεται το αναμενόμενο ωφέλιμο φορτίο. Η χρήση παραδειγμάτων βοηθά τους δυνητικούς χρήστες να χρησιμοποιούν ένα API. Ένα πρότυπο που χρησιμοποιείται για την περιγραφή των APIs είναι το OpenAPI. Επιτρέπει τη χρήση είτε του JSON, είτε του YAML για την περιγραφή των APIs.

Αξιοπιστία (Credibility)

Τα δεδομένα θεωρούνται αξιόπιστα εάν βασίζονται σε αξιόπιστες πηγές. Η αξιοπιστία περιγράφει τον βαθμό στον οποίο τα δεδομένα έχουν χαρακτηριστικά που μπορούν να θεωρηθούν αληθή από τους χρήστες.

Επομένως, αυτός ο δείκτης εξαρτάται σε μεγάλο βαθμό από την αντίληψη του χρήστη. Ωστόσο, η αξιοπιστία των δεδομένων, ενδέχεται να αυξηθεί εάν παρέχονται συγκεκριμένες πληροφορίες, σχετικά με τα συμφραζόμενα, όπως πληροφορίες σχετικά με τον αρχικό εκδότη, το σημείο επαφής και τον κάτοχο του συνόλου των δεδομένων.

Κεφάλαιο 3ο – Το πρότυπο ISO στην ποιότητα δεδομένων

Ενότητα 3.1 – Το πρότυπο ISO25000

Με βάση το πρότυπο ISO, ποιότητα σημαίνει το σύνολο των χαρακτηριστικών μιας οντότητας που επηρεάζουν την ικανότητά της να ικανοποιεί προκαθορισμένες και υπονοούμενες ανάγκες [21]. Τα πρότυπα της σειράς ISO/IEC 25000 αποτελούνται απο τους εξής τομείς:

- Διαχείριση ποιότητας
- Μοντέλα ποιότητας
- Μέτρα ποιότητας
- Αξιολόγηση ποιότητας

Όσον αφορά την αξιολόγηση της ποιότητας των δεδομένων, τα πιο σχετικά πρότυπα είναι τα ISO/IEC 25012 και ISO/IEC 25024. Το ISO/IEC 25012 [26] προσδιορίζει και παρέχει μια ταξινόμηση για τα χαρακτηριστικά που καθορίζουν την ποιότητα των δεδομένων. Όσον αφορά την ταξινόμηση των χαρακτηριστικών ποιότητας δεδομένων, αυτό το πρότυπο τα ταξινομεί από δύο απόψεις:

- Εγγενής ποιότητα δεδομένων, που αναφέρεται στο βαθμό στον οποίο τα χαρακτηριστικά ποιότητας δεδομένων έχουν εγγενείς δυνατότητες να ικανοποιήσουν τις ανάγκες δεδομένων.
- Ποιότητα δεδομένων που εξαρτάται από το σύστημα, που αναφέρεται στο βαθμό στον οποίο η ποιότητα των δεδομένων επιτυγχάνεται, διατηρείται μέσω ενός πληροφοριακού συστήματος και εξαρτάται από το συγκεκριμένο τεχνολογικό πλαίσιο στο οποίο χρησιμοποιούνται τα δεδομένα.

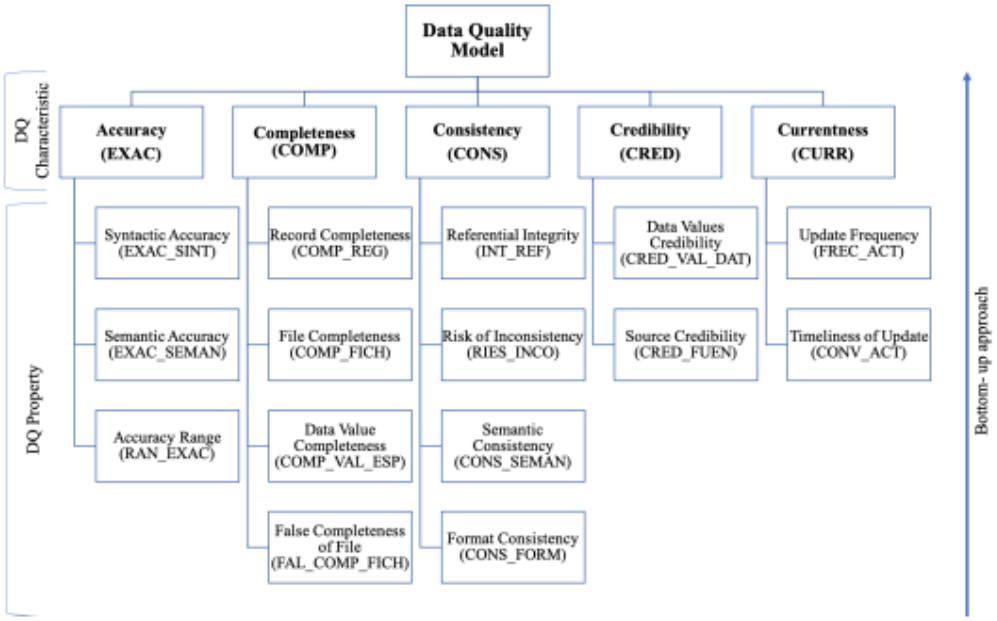
Λόγω του εύρους στην τεχνολογική φύση των δεδομένων, η ανάπτυξη γενικών μέτρων που θα επέτρεπαν τη σύγκριση μεταξύ διαφορετικών οργανισμών καθίσταται σχεδόν αδύνατη, μόνο τα εγγενή χαρακτηριστικά ποιότητας δεδομένων έχουν ληφθεί υπόψη για το περιβάλλον αξιολόγησης ποιότητας δεδομένων. Τα εγγενή χαρακτηριστικά ποιότητας δεδομένων περιγράφονται στον Πίνακα 3.1.

Πίνακας 3.1 Εγγενή χαρακτηριστικά ποιότητας δεδομένων που ορίζονται στο ISO/IEC 25012. (Πηγή: [26])

Χαρακτηριστικό	**Ορισμός**
Ακρίβεια (Accuracy)	Ο βαθμός στον οποίο τα δεδομένα έχουν χαρακτηριστικά που αντιπροσωπεύουν σωστά την πραγματική τιμή/αξία του επιδιωκόμενου χαρακτηριστικού μιας έννοιας ή γεγονότος σε ένα συγκεκριμένο πλαίσιο χρήσης.
Πληρότητα (Completeness)	Ο βαθμός στον οποίο τα δεδομένα του θέματος που σχετίζονται με μια οντότητα έχουν τιμές για όλα τα αναμενόμενα χαρακτηριστικά σε ένα συγκεκριμένο πλαίσιο χρήσης.
Συνοχή (Consistency)	Ο βαθμός στον οποίο τα δεδομένα έχουν χαρακτηριστικά που είναι απαλλαγμένα από αντιφάσεις και είναι συνεπή με άλλα δεδομένα σε ένα συγκεκριμένο πλαίσιο χρήσης.
Αξιοπιστία (Credibility)	Ο βαθμός στον οποίο τα δεδομένα έχουν χαρακτηριστικά που θεωρούνται αληθή και πιστευτά από τους χρήστες σε ένα συγκεκριμένο πλαίσιο χρήσης.
Επικαιρότητα (Currentness)	Ο βαθμός στον οποίο τα δεδομένα έχουν χαρακτηριστικά που συνεπή με τον χρόνο ενδιαφέροντος του συγκεκριμένου πλαισίου χρήσης.

Το ISO/IEC 25024 [27] είναι εκείνο το απόσπασμα της σειράς διεθνών προτύπων SQuaRE που καθιερώνει τη σχέση μεταξύ της έννοιας του χαρακτηριστικού ποιότητας δεδομένων (που εισήχθη με τον ISO/IEC 25012) και της έννοιας της "ιδιότητας ποιότητας". Η ιδιότητα ποιότητας δεδομένων είναι ένα στοιχείο που αντιπροσωπεύει έναν τρόπο αξιολόγησης ορισμένων πτυχών των δεδομένων που περιέχονται σε μια βάση δεδομένων.

Για να αξιολογήσει την ποιότητα μιας βάσης δεδομένων, ένας οργανισμός θα πρέπει να προσδιορίσει τα χαρακτηριστικά ποιότητας δεδομένων και τις αντίστοιχες ιδιότητες ποιότητας δεδομένων που ταιριάζουν καλύτερα στις δηλωμένες απαιτήσεις ποιότητας δεδομένων. Η Εικόνα 3.1 δείχνει μια περίληψη των εγγενών χαρακτηριστικών ποιότητας δεδομένων και των ιδιοτήτων ποιότητας δεδομένων που ορίζονται για κάθε ένα από αυτά.



Εικόνα 3.1 Εγγενή χαρακτηριστικά ποιότητας δεδομένων και σχετικές ιδιότητες ποιότητας δεδομένων κατά ISO/IEC 25012.
(Πηγή: [26]).

Ο Πίνακας 3.2 παρουσιάζει ένα παράδειγμα του τρόπου με τον οποίο περιγράφεται κάθε ιδιότητα ποιότητας δεδομένων κατά ISO/IEC 25024 και τις πληροφορίες που παρέχει σχετικά με τον τρόπο υπολογισμού της τιμής τους. Η ομάδα αξιολόγησης πρέπει να ερμηνεύει πότε οι χαμηλές τιμές για τη μέτρηση μιας ιδιότητας αντιπροσωπεύουν ένα ζήτημα στη βάση δεδομένων.

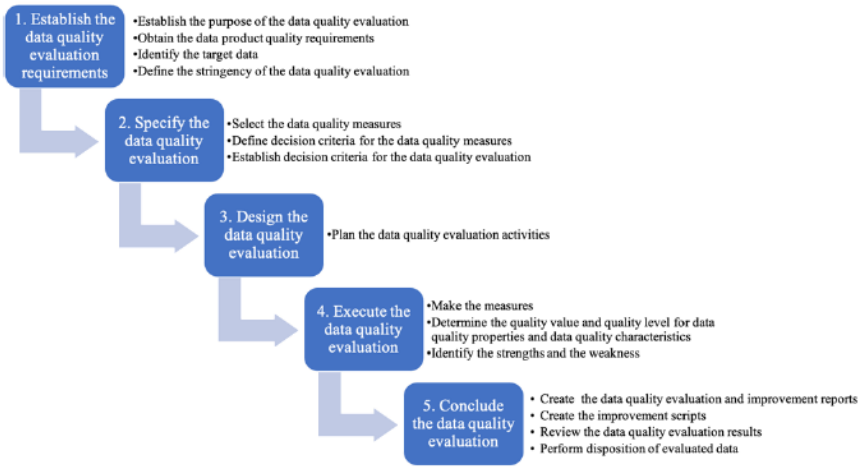
Πίνακας 3.2 Περιγραφή της ιδιότητας "Εύρος Ακρίβειας δεδομένων" (RAN_EXAC), καθώς και η μέτρησή της.
(Πηγή: [27]).

Χαρακτηριστικό ποιότητας δεδομένων	Ακρίβεια
Ιδιότητα ποιότητας δεδομένων	Εύρος Ακρίβειας Δεδομένων
Περιγραφή μέτρησης	Το εύρος ακρίβειας δεδομένων εστιάζει στον έλεγχο, ώστε οι τιμές δεδομένων να περιλαμβάνονται στα απαιτούμενα διαστήματα. Η τιμή λαμβάνεται ως ο λόγος των εγγραφών σε ένα αρχείο δεδομένων του οποίου οι τιμές για τα πεδία βρίσκεται εντός των καθορισμένων διαστημάτων.
Τύπος υπολογισμού	X = A / B A = αριθμός στοιχείων δεδομένων που έχουν μια τιμή που περιλαμβάνεται σε ένα καθορισμένο διάστημα (δηλαδή, κυμαίνεται από το ελάχιστο έως το μέγιστο) B = αριθμός στοιχείων δεδομένων για τα οποία μπορεί να οριστεί ένα απαιτούμενο διάστημα τιμών
Κλίμακα	Αναλογία
Εύρος τιμών	[0.0, 1.0]

Το πρότυπο ISO/IEC 25024 δεν εξετάζει συγκεκριμένα τον τρόπο με τον οποίο θα πρέπει να συγκεντρωθούν τα μέτρα που αντιστοιχούν στις ιδιότητες ποιότητας δεδομένων, για τον υπολογισμό του συνολικού επιπέδου ποιότητας για κάθε χαρακτηριστικό ποιότητας δεδομένων. Εναπόκειται στον οργανισμό που διενεργεί την αξιολόγηση της ποιότητας των δεδομένων, να προσδιορίσει επαρκώς τον τρόπο συγκέντρωσης των μετρήσεων για τις ιδιότητες ποιότητας δεδομένων, προκειμένου να αποκτήσει την τιμή επιπέδου ποιότητας για τα χαρακτηριστικά ποιότητας δεδομένων.

Ενότητα 3.2 – Πρότυπο ISO25000 & αξιολόγηση ποιότητας δεδομένων

Η διαδικασία αξιολόγησης, η οποία περιλαμβάνει τις δραστηριότητες, τις εισροές και τις εκροές που απαιτούνται για τη διεξαγωγή της αξιολόγησης της ποιότητας δεδομένων (Εικόνα 3.2), αποτελείται από πέντε συστηματικές, αυστηρές και επαναλαμβανόμενες δραστηριότητες μέσω των οποίων λαμβάνονται τα μέτρα για τα επιλεγμένα χαρακτηριστικά ποιότητας των δεδομένων. Οι σχετικές πτυχές για κάθε δραστηριότητα της διαδικασίας αξιολόγησης ποιότητας δεδομένων περιγράφονται παρακάτω:



Εικόνα 3.2 Δραστηριότητες αξιολόγησης ποιότητας δεδομένων κατά ISO/IEC 25024.
(Πηγή: [37]).

Στη Δραστηριότητα 1 της παραπάνω διαδικασίας, πραγματοποιείται καθορισμός των απαιτήσεων αξιολόγησης ποιότητας δεδομένων, μεταξύ της ομάδας αξιολόγησης του

διαπιστευμένου εργαστηρίου και του προσωπικού του οργανισμού που επιδιώκει να αξιολογήσει το επίπεδο ποιότητας δεδομένων μιας βάσης δεδομένων. Η ομάδα αξιολόγησης προτείνει το μοντέλο και τα μέτρα ποιότητας δεδομένων στο προσωπικό του οργανισμού, με στόχο την επιλογή εκείνων των ποιοτικών χαρακτηριστικών που ταιριάζουν καλύτερα στις απαιτήσεις της ποιότητας δεδομένων και ο οργανισμός αποφασίζει να επιλέξει όλα ή μόνο ένα υποσύνολο των χαρακτηριστικών ποιότητας δεδομένων. Από την πλευρά του διαπιστευμένου εργαστηρίου, είναι σημαντικό ο οργανισμός να ορίσει πρώτα τους επιχειρηματικούς κανόνες (παράδειγμα αυτών στον Πίνακα 3.3) για τη βάση δεδομένων του με λεπτομέρεια, καθώς αυτοί οι επιχειρηματικοί κανόνες αποτελούν τη βάση για τη διαδικασία αξιολόγησης. Με αυτόν τον τρόπο, το έγγραφο που περιέχει όλους τους επιχειρηματικούς κανόνες πρέπει να ελεγχθεί από τα ενδιαφερόμενα μέρη και την ομάδα αξιολόγησης μέχρι να επιτευχθεί συμφωνία για την επάρκειά του. Κατόπιν τούτου, το προσωπικό του οργανισμού με την υποστήριξη της ομάδας αξιολόγησης πρέπει να κατηγοριοποιήσει τους επιχειρηματικούς κανόνες στις ιδιότητες ποιότητας δεδομένων του μοντέλου αξιολόγησης ποιότητας, λαμβάνοντας υπόψη τη φύση των επιχειρηματικών κανόνων και τον σκοπό των ιδιοτήτων ποιότητας δεδομένων.

Πίνακας 3.3 Παράδειγμα επιχειρηματικών κανόνων για βάση δεδομένων.
(Πηγή: [37]).

Ταυτότητα Επιχειρηματικού κανόνα	**Παράδειγμα επιχειρηματικού κανόνα**
1	person.id πρέπει αντιπροσωπέυεται από 9 χαρακτήρες, από τους οποίους τα πρώτα 8 ψηφία πρέπει να είναι στο διάστημα [0-9] και τα τελευταίο ψηφίο πρέπει να είναι χαρακτήρας γράμματος [A-Z].
2	person.ipaddress πρέπει να εκφράζεται από 4 αριθμούς στο διάστημα [0-255], χωριζόμενα από μια τελεία, για παράδειγμα 126.12.4.89.
3	warning.type λαμβάνει μια από τις ακόλουθες τιμές: {IT GENERAL, SUPERCOMPUTATION, HR}.

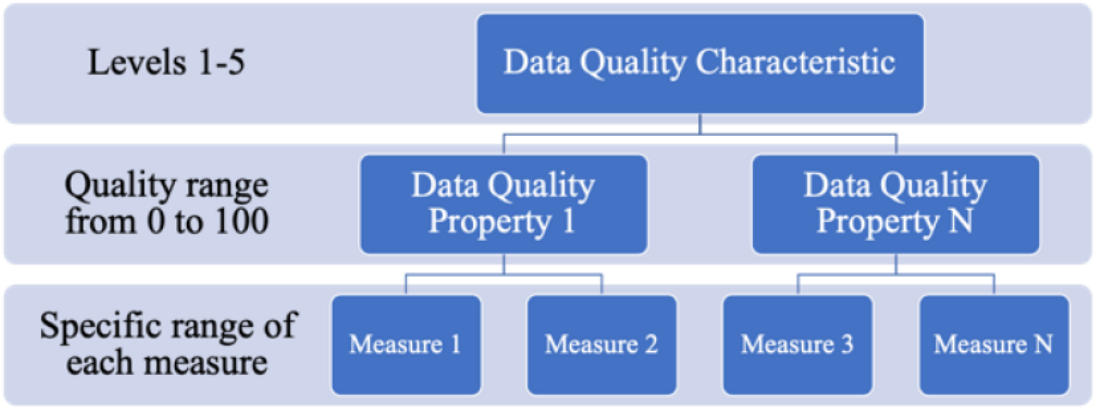
Στη Δραστηριότητα 2 της παραπάνω διαδικασίας, πραγματοποιείται καθορισμός της αξιολόγησης της ποιότητας των δεδομένων, καθώς σε αυτή τη δραστηριότητα παρουσιάζονται οι μέθοδοι μέτρησης και τα αντίστοιχα επίπεδα ποιότητας για τα μέτρα κάθε ιδιότητας ποιότητας δεδομένων, έτσι ώστε ο οργανισμός να μπορεί να κατανοήσει τα αποτελέσματα της αξιολόγησης. Με αυτό τον τρόπο γίνεται δυνατή η βελτίωση των πληροφοριών που παρέχονται καθώς και ο συντονισμός ορισμένων λεπτομερειών

εφαρμογής των επιχειρηματικών κανόνων για το χώρο αποθήκευσης δεδομένων. Αυτή η γνώση θα χρησιμοποιηθεί από την ομάδα αξιολόγησης κατά τη διεξαγωγή της Δραστηριότητας 4 προκειμένου να αναπτυχθούν τα σενάρια αξιολόγησης.

Τα επίπεδα ποιότητας και οι μέθοδοι μέτρησης (καθώς και τα υποκείμενα στοιχεία, όπως οι τιμές κατωφλίου που χρησιμοποιούνται στις αντίστοιχες μετρήσεις) καθορίζονται από το διαπιστευμένο κατά ISO εργαστήριο ως μέρος της προσαρμογής του μοντέλου ποιότητας δεδομένων, το οποίο υποστηρίζεται από διάφορες πιλοτικές εμπειρίες. Για παράδειγμα, οι τιμές κατωφλίου που χρησιμοποιούνται στον υπολογισμό των επιπέδων ποιότητας δεδομένων επιλέγονται, ώστε να αντιπροσωπεύουν πόσο καλά μπορεί να αποδώσει ο οργανισμός σε διαφορετικά σενάρια, όπου τα ανεπαρκή επίπεδα ποιότητας δεδομένων είναι η πηγή διαφορετικών προβλημάτων. Όσο υψηλότερο είναι το επίπεδο ποιότητας (και η αντίστοιχη τιμή κατωφλίου), τόσο λιγότερο πιθανό αντίκτυπο μπορεί να έχει το ζήτημα της ποιότητας των δεδομένων στον οργανισμό. Ως εκ τούτου, προκύπτει ως αποτέλεσμα η κατανόηση των λεπτομερειών της διαδικασίας αξιολόγησης ποιότητας δεδομένων από τον οργανισμό.

Κατά τη Δραστηριότητα 3 πραγματοποιείται ο σχεδιασμός της αξιολόγησης της ποιότητας των δεδομένων, ενώ το εύρος της αξιολόγησης καθορίζεται λεπτομερώς. Επιπλέον, καταρτίζεται πλήρως το σχέδιο αξιολόγησης ποιότητας δεδομένων που περιέχει τις δραστηριότητες και τα καθήκοντα που απαιτούνται για τη διεξαγωγή της αξιολόγησης. Ως αποτέλεσμα λαμβάνεται το σχέδιο αξιολόγησης ποιότητας δεδομένων, το οποίο περιέχει τις συγκεκριμένες λεπτομέρειες (π.χ. χρονοδιάγραμμα, εργασίες, πόρους κ.λπ.) για τη διεξαγωγή της αξιολόγησης ποιότητας δεδομένων στις οντότητες δεδομένων που περιλαμβάνονται στο πεδίο εφαρμογής.

Κατά τη Δραστηριότητα 4 πραγματοποιείται η αξιολόγηση ποιότητας δεδομένων, ενώ το επίπεδο ποιότητας για τα επιλεγμένα χαρακτηριστικά ποιότητας δεδομένων επιτυγχάνεται, ακολουθώντας το σχέδιο αξιολόγησης που καθορίστηκε κατά τις Δραστηριότητες 1 έως 3. Αυτές οι τιμές επιπέδου ποιότητας για τα χαρακτηριστικά ποιότητας λαμβάνονται ακολουθώντας συγκεκριμένη ιεραρχία όπως φαίνεται στην Εικόνα 3.3.



Εικόνα 3.3 Ιεραρχία των στοιχείων που συνιστούν το μοντέλο ποιότητας δεδομένων για την αξιολόγηση της ποιότητας των δεδομένων.
(Πηγή: [37]).

Στην υπόψη Δραστηριότητα περιλαμβάνεται η χρήση ορισμένων εννοιών (π.χ. επίπεδα ποιότητας, εύρη ποιότητας και συναρτήσεις προφίλ) που εξηγούνται λεπτομερώς στο M. Rodríguez et al. (2016) [28]. Οι εργασίες που απαιτούνται κατά την εκτέλεση της δραστηριότητας είναι οι εξής:

1. Σχεδιασμός σεναρίων αξιολόγησης.
2. Εκτέλεση των σεναρίων αξιολόγησης.
3. Παραγωγή της τιμής ποιότητας για τις ιδιότητες ποιότητας δεδομένων.
4. Εξαγωγή του επιπέδου ποιότητας για τις ιδιότητες ποιότητας δεδομένων από την ποιοτική τους αξία.
5. Προσδιορισμός του επιπέδου ποιότητας για τα επιλεγμένα δεδομένα ποιοτικά χαρακτηριστικά.

Ως αποτέλεσμα αυτής της δραστηριότητας, προκύπτουν τα ακόλουθα αποτελέσματα:

- Αξία και επίπεδο ποιότητας για κάθε ιδιότητα ποιότητας δεδομένων που αξιολογείται.
- Τιμή ποιότητας για κάθε χαρακτηριστικό ποιότητας δεδομένων που έχει επιλεγεί για την αξιολόγηση.
- Μειονεκτήματα και πλεονεκτήματα που εντοπίστηκαν για κάθε ιδιότητα ποιότητας δεδομένων.

Κατά τη Δραστηριότητα 5 πραγματοποιείται η ολοκλήρωση της αξιολόγησης της ποιότητας των δεδομένων, καθώς στοιχειοθετείται λεπτομερής έκθεση αξιολόγησης, η οποία αντικατοπτρίζει τα επίπεδα ποιότητας που επιτυγχάνονται για τα επιλεγμένα

ποιοτικά χαρακτηριστικά δεδομένων, καθώς και τις τιμές που λαμβάνονται για τις αντίστοιχες ιδιότητες ποιότητας δεδομένων. Επιπλέον, μπορεί να παραχθεί και να παρασχεθεί στον οργανισμό μια ολοκληρωμένη αναφορά βελτίωσης, η οποία περιγράφει λεπτομερώς τις αδυναμίες και τα δυνατά σημεία που σχετίζονται με κάθε ιδιότητα ποιότητας δεδομένων. Αυτή η αναφορά βελτίωσης εστιάζει σε εκείνες τις ιδιότητες που δεν πέτυχαν επαρκές επίπεδο ποιότητας και παρουσιάζει λεπτομέρειες σχετικά με τις αιτίες του ζητήματος, έτσι ώστε ο οργανισμός να αναλάβει δράση για τη βελτίωσή τους.

Απο τα παραπάνω προκύπτει ότι η προτυποποίηση της ποιότητας δεδομένων κατά ISO, απαιτεί ενδελεχή ανάλυση της ποιότητας δεδομένων και των διαδικασιών που οδηγούν σε αυτή.

Κεφάλαιο 4ο – Αλγόριθμοι, μέθοδοι & τεχνικές ανάλυσης ποιότητας δεδομένων

Ενότητα 4.1 – Εισαγωγή

Ανάπτυξη έχει παρουσιαστεί τα τελευταία χρόνια στον τομέα της διαχείρισης ποιότητας δεδομένων τόσο σε επίπεδο διαδικασίας όσο και σε επίπεδο προϊόντος. Αυτό υποστηρίζεται, επίσης, από το γεγονός ότι έχει αναπτυχθεί μια σειρά από πρότυπα, τα οποία αντιμετωπίζουν συγκεκριμένα ζητήματα, χρησιμοποιώντας μοντέλα ποιότητας δεδομένων (π.χ. ISO 25000), μοντέλα ωριμότητας διεργασιών (π.χ. ISO 15504, CMMI), καθώς και πρότυπα που επικεντρώνται κυρίως στην επαλήθευση και επικύρωση λογισμικού (ISO 12207, IEEE 1028, κ.α.). Αυτά τα πρότυπα λαμβάνονται υπόψη παγκοσμίως για περίπου δεκαπέντε χρόνια [29].

Ενότητα 4.2 – Στρατηγικές και τεχνικές στην ποιότητα δεδομένων

Στην προσπάθεια αναβάθμισης της ποιότητας των δεδομένων, οι μεθοδολογίες υιοθετούν δύο γενικούς τύπους στρατηγικών, δηλαδή βάσει δεδομένων (data-driven) και βάσει διεργασιών (process-driven). Οι στρατηγικές που εστιάζουν στα δεδομένα βελτιώνουν την ποιότητα των δεδομένων, τροποποιώντας άμεσα την αξία των δεδομένων. Για παράδειγμα, παλιές τιμές δεδομένων ενημερώνονται ανανεώνοντας μια βάση δεδομένων με δεδομένα από μια πιο πρόσφατη βάση δεδομένων. Οι στρατηγικές που εστιάζουν στις διαδικασίες βελτιώνουν την ποιότητα επανασχεδιάζοντας τις διαδικασίες που δημιουργούν ή τροποποιούν δεδομένα. Για παράδειγμα, μια διαδικασία μπορεί να επανασχεδιαστεί χρησιμοποιώντας μια δραστηριότητα που ελέγχει τη μορφή των δεδομένων, πριν από την αποθήκευση.

Οι άνωθεν στρατηγικές εφαρμόζουν μια ποικιλία τεχνικών: αλγόριθμους, ευρετικές μεθόδους και δραστηριότητες που βασίζονται στη γνώση. Οι στρατηγικές που εστιάζουν στα δεδομένα εφαρμόζουν κατά βάση τις εξής τεχνικές:

- Η εύρεση και χρήση νέων δεδομένων, η οποία βελτιώνει τα δεδομένα με την χρήση δεδομένων υψηλότερης ποιότητας για να αντικαταστήσει τις τιμές που οδηγούν σε προβλήματα ποιότητας.

- Η τυποποίηση (ή κανονικοποίηση) δεδομένων, βάσει της οποίας αντικαθίστανται ή συμπληρώνονται μη τυπικές τιμές δεδομένων με αντίστοιχες τιμές που συμμορφώνονται με το εκάστοτε πρότυπο, ώστε να υπάρχει ομοιομορφία στη βάση δεδομένων.
- Η σύνδεση εγγραφών, βάσει της οποίας γίνεται προσδιορισμός μιας αναπαράστασης δεδομένων, η οποία μπορεί να παρουσιάζεται σε αρκετούς πίνακες και να αναφέρεται στο ίδιο αντικείμενο του πραγματικού κόσμου.
- Η ενοποίηση δεδομένων και σχημάτων, βάσει της οποίας διαμορφώνεται μια ομοιογενής προβολή των δεδομένων που αποκτώνται από ετερογενείς πηγές δεδομένων. Στα κατανεμημένα, συνεργατικά και τα συστήματα πληροφοριών Peer-to-peer (P2P), οι πηγές δεδομένων χαρακτηρίζονται από διάφορα είδη ετερογένειας που μπορούν γενικά να ταξινομηθούν σε τεχνολογικές ετερογένειες, ετερογένειες σχήματος και ετερογένειες σε επίπεδο στιγμιότυπου.
    - Οι τεχνολογικές ετερογένειες οφείλονται στη χρήση προϊόντων από διαφορετικούς προμηθευτές, που χρησιμοποιούνται σε διάφορα επίπεδα μιας υποδομής πληροφοριών και επικοινωνιών.
    - Οι ετερογένειες σχήματος προκαλούνται κυρίως από τη χρήση διαφορετικών μοντέλων δεδομένων, όπως στην περίπτωση μιας πηγής που υιοθετεί το μοντέλο σχεσιακών δεδομένων (RM) και μιας διαφορετικής πηγής που υιοθετεί το μοντέλο δεδομένων XML ή/και διαφορετικών αναπαραστάσεων για το ίδιο αντικείμενο, όπως δύο σχεσιακές πηγές που αντιπροσωπεύουν ένα αντικείμενο ως πίνακα και ως χαρακτηριστικό.
    - Οι ετερογένειες σε επίπεδο στιγμιότυπου προκαλούνται από διαφορετικές, αντικρουόμενες τιμές δεδομένων που παρέχονται από διαφορετικές πηγές για τα ίδια αντικείμενα. Για παράδειγμα, αυτός ο τύπος ετερογένειας μπορεί να προκληθεί από ανεξάρτητες και ανεπαρκώς συντονισμένες διαδικασίες που τροφοδοτούν τις διαφορετικές πηγές δεδομένων.

  Η ενοποίηση δεδομένων οφείλει να αντιμετωπίζει όλους τους τύπους αυτών των ετερογενειών.
- Η αξιοπιστία της πηγής, βάσει της οποίας επιλέγονται πηγές δεδομένων με βάση την ποιότητα των δεδομένων τους.

- Ο εντοπισμός και η διόρθωση σφαλμάτων, τα οποία εντοπίζουν και εξαλείφουν σφάλματα ποιότητας δεδομένων εντοπίζοντας τις εγγραφές που δεν πληρούν ένα δεδομένο σύνολο κανόνων ποιότητας. Αυτές οι τεχνικές μελετώνται κυρίως στον στατιστικό τομέα. Σε σύγκριση με τα μεμονωμένα δεδομένα, τα συγκεντρωτικά στατιστικά δεδομένα, όπως ο μέσος όρος, το άθροισμα ή το μέγιστο μεταξύ τιμών είναι λιγότερο ευαίσθητα σε πιθανώς λανθασμένο εντοπισμό και διόρθωση τιμών. Τεχνικές εντοπισμού και διόρθωσης σφαλμάτων έχουν προταθεί για ασυνέπειες, ελλιπή δεδομένα και ακραίες τιμές.
- Η βελτιστοποίηση κόστους βάσει της οποίας ορίζονται ενέργειες βελτίωσης της ποιότητας μέσω ενός συνόλου διαστάσεων, ελαχιστοποιώντας το κόστος.

Οι στρατηγικές που εστιάζουν στις διαδικασίες εφαρμόζουν κατά βάση τις εξής δύο τεχνικές:

- Ο έλεγχος διαδικασίας εισάγει ελέγχους και διαδικασίες ελέγχου στη διαδικασία παραγωγής δεδομένων όταν: (1) δημιουργούνται νέα δεδομένα, (2) ενημερώνονται σύνολα δεδομένων ή (3) γίνεται πρόσβαση σε νέα σύνολα δεδομένων από τη διαδικασία. Εφαρμόζεται, λοιπόν, μια στρατηγική σε γεγονότα τροποποίησης δεδομένων, αποφεύγοντας έτσι την υποβάθμιση των δεδομένων και τη διάδοση σφαλμάτων.
- Ο επανασχεδιασμός της διαδικασίας επανασχεδιάζει τις διαδικασίες προκειμένου να αρθούν οι αιτίες της κακής ποιότητας δεδομένων και εισάγει νέες δραστηριότητες που παράγουν δεδομένα υψηλότερης ποιότητας. Εάν ο επανασχεδιασμός της διαδικασίας είναι ριζικός, αυτή η τεχνική αναφέρεται ως ανασχεδιασμός επιχειρησιακών διαδικασιών [30,31].

Στη βιβλιογραφία, έχουν γίνει αρκετές συγκρίσεις μεταξύ των τεχνικών και της αποτελεσματικότητάς τους σε συνάρτηση με το επωμιζόμενο κόστος. Γενικά, μακροπρόθεσμα, οι τεχνικές που βασίζονται στη διαδικασία διαπιστώνεται ότι υπερτερούν των τεχνικών που βασίζονται σε δεδομένα, καθώς εξαλείφουν τις βασικές αιτίες των προβλημάτων ποιότητας. Ωστόσο, υπό βραχυπρόθεσμη προοπτική, ο επανασχεδιασμός της διαδικασίας μπορεί να είναι εξαιρετικά δαπανηρός [32,33]. Αντίθετα, οι στρατηγικές που βασίζονται σε δεδομένα αναφέρεται ότι είναι οικονομικά αποδοτικές βραχυπρόθεσμα, αλλά ακριβές μακροπρόθεσμα. Είναι κατάλληλες για μεμονωμένες εφαρμογές και ως εκ τούτου, συνιστώνται για μη δυναμικά δεδομένα.

Ενότητα 4.3 – Μέθοδοι Ανάλυσης και Ποιότητας Δεδομένων

Οι Liebchen και Shepperd (2008) [34] έκαναν συστηματική ανασκόπηση των εμπειρικών μελετών λογισμικού με στόχο να βρεθούν μελέτες, οι οποίες λαμβάνουν υπόψη, τουλάχιστον εν μέρει, την ποιότητα των δεδομένων. Διαπίστωσαν ότι από τις περισσότερες από 500 μελέτες για τη μηχανική λογισμικού που είχαν εντοπίσει, μόνο 23 αναφέρονταν στην ποιότητα των δεδομένων. Για τους πιο δημοφιλείς τρόπους αντιμετώπισης της ποιότητας δεδομένων, οι χειροκίνητες τεχνικές βρέθηκαν να είναι οι πιο δημοφιλείς (σχεδόν το 50% των μελετών) και μερικές μελέτες ανέφεραν τεχνικές πρόληψης, όπως εργαλεία υποστήριξης για τη συλλογή δεδομένων που είναι κυρίως εξωτερικά, π.χ. μη ενσωματωμένα στο σχεδιασμό του πληροφοριακού συστήματος.

Οι D.C. Corrales et al. [35] στο πλαίσιο των πληροφοριακών συστημάτων, εξέτασαν περισσότερες από 100 εργασίες για να εντοπίσουν τα κύρια ζητήματα ποιότητας δεδομένων. Ανάλογα με το εκάστοτε ζήτημα της ποιότητας των δεδομένων που εξετάζεται στο φάσμα του αγροτικού τομέα, π.χ. την ετερογένεια δεδομένων, τις ακραίες τιμές, την επικαιρότητα, την ακεραιότητα, την ασυνέπεια, τον θόρυβο, τον πλεονασμό και την ποσότητα των δεδομένων, οι βασικές λύσεις ήταν κυρίως οι παρακάτω, όπως συμβαίνει συνήθως σε κάθε τομέα:

- η μη εποπτευόμενη μάθηση (π.χ. αλγόριθμοι τμηματικών συστάδων όπως k-means, σταθμισμένοι k-means, πυκνοτικοί και ιεραρχικοί αλγόριθμοι και συνδυασμός επιμερισμένων και ασαφών αλγορίθμων),
- η εποπτευόμενη μάθηση (π.χ. Decision trees, Linear regression, η μέθοδος Support Vector Machines (SVM), τα νευρωνικά δίκτυα, το Bayesian δίκτυο),
- οι στατιστικές μέθοδοι.

Σε άλλη μελέτη, οι Batini et al. [36] παρουσιάζουν μια συγκριτική αναφορά 13 μεθοδολογιών ποιότητας δεδομένων, οποίες είναι οι εξής:

- Total Data Quality Management (TDQM)
- Data Warehouse Quality Methodology (DWQ)
- Total Information Quality Management (TIQM)
- Methodology for information Quality Assessment (AIMQ)
- Canadian Institute for Health Information methodology (CIHI)

- Data Quality Assessment (DQA)
- Information Quality Measurement (IQM)
- ISTAT (developed by Italian National Bureau of Census)
- Activity-based Measuring and Evaluating of product information Quality (AMEQ)
- Cost-effect Of Low Data Quality (COLDQ)
- Data Quality in Cooperative Information Systems (DaQuinCIS)
- Methodology for the Quality Assessment of Financial Data (QAFD)
- Comprehensive methodology for Data Quality management (CDQ)

Αυτές οι μεθοδολογίες αξιολογήθηκαν σε διάφορες πτυχές τους, όπως μεθοδολογικές φάσεις και βήματα, στρατηγικές και τεχνικές, διαστάσεις ποιότητας δεδομένων, τύπος δεδομένων και τύποι πληροφοριακών συστημάτων που εξετάζονται από κάθε μεθοδολογία. Διαπίστωσαν ότι μόνο 5 από τις 13 μεθοδολογίες αναφέρονται σε ανάλυση απαιτήσεων ποιότητας δεδομένων, ενώ 6 από τις 13 αναφέρονται σε μοντελοποίηση διαδικασίας, με μόνο 3 από αυτές να σχετίζονται και με τις δύο. Οι ομοιότητες και οι διαφορές σε αυτές τις μεθοδολογίες οδήγησαν τους συγγραφείς να τις ταξινομήσουν στις τέσσερις παρακάτω κατηγορίες:

1. πλήρεις μεθοδολογίες που παρέχουν υποστήριξη στις φάσεις αξιολόγησης και βελτίωσης και αντιμετωπίζουν τόσο τεχνικά όσο και οικονομικά ζητήματα,
2. μεθοδολογίες ελέγχου που επικεντρώνονται στη φάση αξιολόγησης και πολύ περιορισμένη εστίαση (ή καθόλου) στη φάση βελτίωσης,
3. επιχειρησιακές μεθοδολογίες που επικεντρώνονται στα τεχνικά ζητήματα τόσο της φάσης αξιολόγησης όσο και της φάσης βελτίωσης, αλλά δεν εστιάζουν στα οικονομικά ζητήματα,
4. οικονομικές μεθοδολογίες που εστιάζουν στην εκτίμηση κόστους.

Σημαντικές αλλαγές επέφερε το πρότυπο ISO για την ποιότητα των δεδομένων (ISO/IEC 25012). Τα πρότυπα ISO είναι ένας κοινός τρόπος επίτευξης συμφωνίας σε ένα δεδομένο θέμα, με ένα και ενιαίο σύνολο απαιτήσεων και κατευθυντήριων γραμμών που δεν διαφέρουν σημαντικά από τη μια περίπτωση στην άλλη, απαιτώντας μόνο προσαρμογή και όχι νέους ορισμούς. Αυτό είναι ιδιαίτερα σημαντικό για τον όρο ποιότητας δεδομένων και τις διαστάσεις ποιότητας δεδομένων. Επιπλέον, είναι γνωστό ότι αυτό το πρότυπο χρησιμοποιείται ευρέως στην πράξη για την επαλήθευση και τη

διατήρηση της ποιότητας δεδομένων για ένα σύστημα εν λειτουργία. Όπως αναλύθηκε παραπάνω, οι Gualo et al. (2021) [37] διαπίστωσαν ότι αυτό θα μπορούσε να είναι μια πρόκληση για τα ενδιαφερόμενα μέρη που εμπλέκονται στη διαδικασία, επειδή:

1. ορισμένοι από αυτούς δεν γνωρίζουν πραγματικά την ποιότητα των δεδομένων, ακόμη και όταν χρησιμοποιείται το πρότυπο,
2. οι περισσότεροι από αυτούς δεν μπορούν να ορίσουν απαιτήσεις ποιότητας δεδομένων, επιχειρηματικούς κανόνες ή πιθανά ζητήματα ή τα περιορίζουν απλώς στην πληρότητα και την επικαιρότητα,
3. άλλοι χρησιμοποιούν ad-hoc προσεγγίσεις που σχετίζονται κυρίως με τις ιδιαίτερες περιπτώσεις χρήσης τους, δηλαδή δεν είναι γενικεύσιμες ή επαρκώς λεπτομερείς για να τις αναπαράγουν.

Ενότητα 4.3.1 – Μέθοδος Total Data Quality Management (TDQM)

Η μεθοδολογία Total Data Quality Management (TDQM) είναι η πρώτη γενική μεθοδολογία που δημοσιεύτηκε στη βιβλιογραφία για την ποιότητα των δεδομένων [38]. Η μέθοδος TDQM έχει χρησιμοποιηθεί σε μεγάλο βαθμό για πρωτοβουλίες ανασυγκρότησης των δεδομένων διαφόρων οργανισμών. Βασικός στόχος της μεθόδου είναι να επεκτείνει στην ποιότητα των δεδομένων, τις αρχές της Διαχείρισης Ολικής Ποιότητας (Total Quality Management / TQM) [39]. Στη διαχείριση λειτουργιών, η μέθοδος προσφέρει μεθοδολογικές κατευθυντήριες γραμμές που στοχεύουν στην εξάλειψη των διαφορών μεταξύ των αποτελεσμάτων των λειτουργικών διαδικασιών και των απαιτήσεων των πελατών. Λαμβάνοντας υπόψη τις απαιτήσεις, ο ανασχεδιασμός ξεκινά από τη μοντελοποίηση των λειτουργικών διαδικασιών. Ως εκ τούτου, η μέθοδος TDQM προτείνει μια γλώσσα για την περιγραφή των διαδικασιών παραγωγής πληροφοριών (IP), που ονομάζεται IP-MAP [40]. Η γλώσσα IP-MAP έχει επεκταθεί ποικιλοτρόπως, προς την UML και επίσης για την υποστήριξη του οργανωτικού σχεδιασμού.

Σκοπός της μεθόδου TDQM είναι η υποστήριξη βελτίωσης της ποιότητας σε όλο το μήκος της διαδικασίας, από την ανάλυση απαιτήσεων έως και την υλοποίηση. Όπως φαίνεται και στην Εικόνα 4.1, η μέθοδος αποτελείται από τέσσερις φάσεις που

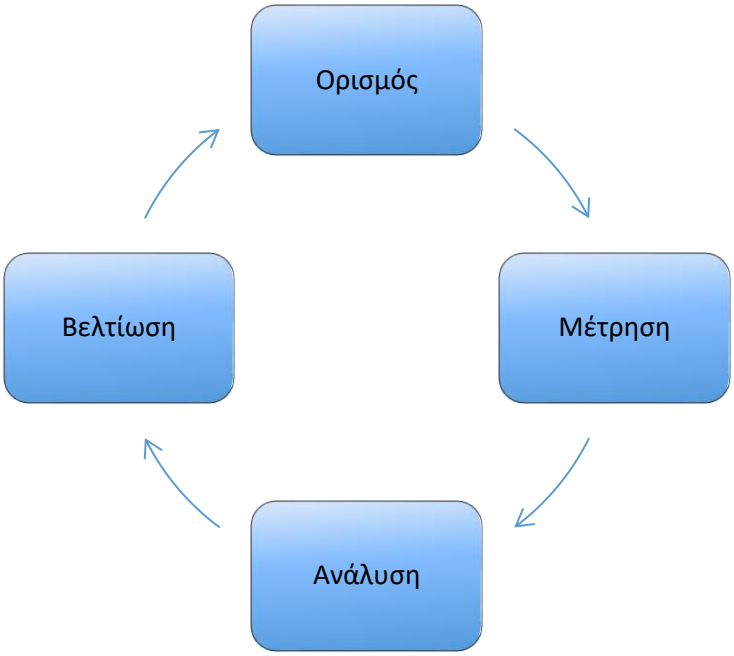
εφαρμόζουν μια διαδικασία συνεχούς βελτίωσης της ποιότητας: ορισμός, μέτρηση, ανάλυση και βελτίωση.

Διακρίνονται τέσσερις ρόλοι που είναι αρμόδιοι για τις διάφορες φάσεις της διαδικασίας βελτίωσης της ποιότητας:

- Προμηθευτές πληροφοριών, οι οποίοι δημιουργούν ή συλλέγουν δεδομένα για την IP.
- Κατασκευαστές πληροφοριών, που σχεδιάζουν, αναπτύσσουν ή διατηρούν δεδομένα.
- Καταναλωτές πληροφοριών, που χρησιμοποιούν δεδομένα στην εργασία τους.
- Διαχειριστές πληροφοριακών διεργασιών, οι οποίοι είναι υπεύθυνοι για τη διαχείριση της διαδικασίας παραγωγής πληροφοριών σε όλη τη διάρκεια του κύκλου ζωής των πληροφοριών.

Η μέθοδος TDQM παρέχει κατευθυντήριες γραμμές για τον τρόπο εφαρμογής της μεθοδολογίας. Κατά την εφαρμογή του TDQM, ένας οργανισμός οφείλει:

- να κατανοεί σαφώς της IP,
- να δημιουργήσει μια ομάδα IP που αποτελείται από ένα ανώτερο στέλεχος ως μάνατζερ διαχείρισης της TDQM («πρωταθλητής», όπως αναφέρεται χαρακτηριστικά), έναν μηχανικό IP που είναι εξοικειωμένος με τη μεθοδολογία TDQM και μέλη οι οποίοι είναι προμηθευτές πληροφοριών, κατασκευαστές και διαχειριστές IP,
- να διδάξει την αξιολόγηση και τη διαχείριση της νοημοσύνης (IQ) σε όλες τις IP,
- να συμβάλλει στη συνεχή βελτίωση της IP.



Εικόνα 4.1 Φάσεις της μεθόδου TDQM.
(Πηγή: [36]).

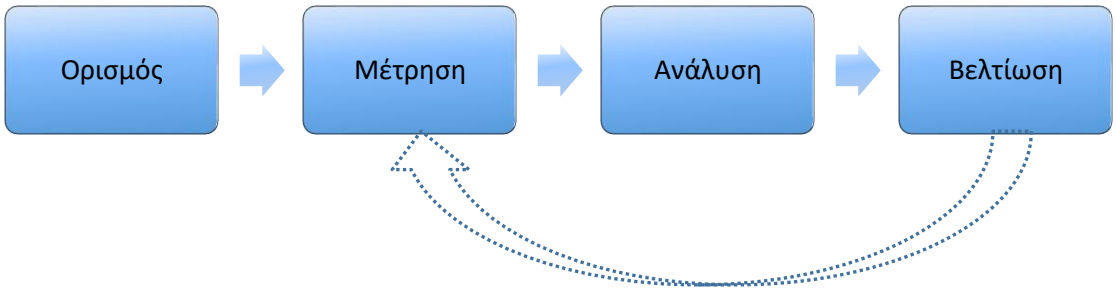
Ενότητα 4.3.2 – Μέθοδος Data Warehouse Quality (DWQ)

Η μέθοδος Data Warehouse Quality (DWQ) αναπτύχθηκε στο πλαίσιο του έργου European Data Warehouse Quality [41]. Εξετάζει τη σχέση μεταξύ στόχων ποιότητας και επιλογών σχεδιασμού σε μια αποθήκη δεδομένων. Οι αποθήκες δεδομένων είναι πολύπλοκα συστήματα που επιβάλλεται να παρέχουν εξαιρετικά συγκεντρωτικά, υψηλής ποιότητας δεδομένα από ετερογενείς πηγές στους αποφασίζοντες. Λόγω της δυναμικής αλλαγής στις απαιτήσεις και το περιβάλλον, τα συστήματα αποθήκης δεδομένων βασίζονται σε μετα-βάσεις δεδομένων (meta databases) για τον έλεγχο της λειτουργίας τους και για την υποβοήθηση της εξέλιξής τους.

Η μεθοδολογία εξετάζει την υποκειμενικότητα της έννοιας της ποιότητας και παρέχει μια ταξινόμηση των στόχων ποιότητας, σύμφωνα με την ομάδα ενδιαφερομένων που θέτει τους στόχους. Ιδιαίτερη σημασία απόδίδεται στην εξέταση της ποικιλίας των στόχων ποιότητας ορίζοντας αντίστοιχα μεταδεδομένα. Πιο συγκεκριμένα, η μεθοδολογία DWQ δηλώνει ότι τα μεταδεδομένα της αποθήκης δεδομένων πρέπει να αντιπροσωπεύουν τρεις προοπτικές:

- μια εννοιολογική επιχειρηματική προοπτική που εστιάζει στο επιχειρηματικό μοντέλο,
- μια λογική προοπτική που εστιάζει στο σχήμα/μορφή της αποθήκης δεδομένων,
- μια φυσική προοπτική που αντιπροσωπεύει το φυσικό επίπεδο μεταφοράς δεδομένων.

Οι άνωθεν προοπτικές αντιστοιχούν στα τρία παραδοσιακά επίπεδα αποθήκευσης δεδομένων, δηλαδή τις πηγές, την αποθήκη δεδομένων και τους πελάτες. Η μέθοδος συσχετίζει κάθε προοπτική με μια άποψη μεταδεδομένων που ονομάζεται Μέτρηση Ποιότητας (Quality Measurement). Από την άποψη της ποιότητας των δεδομένων, τέσσερις κύριες φάσεις χαρακτηρίζουν τη μεθοδολογία: Ορισμός (Definition), Μέτρηση (Measurement), Ανάλυση (Analysis) και Βελτίωση (Improvement) (Εικόνα 4.2).



Εικόνα 4.2 Φάσεις της μεθόδου DWQ.
(Πηγή: [36]).

Ένα από τα πλεονεκτήματα της μεθόδου είναι ότι παρέχει ταξινόμηση των διαστάσεων δεδομένων και ποιότητας λογισμικού στο πλαίσιο της αποθήκης δεδομένων. Στα πλαίσια της μεθόδου ορίζονται τρεις κατηγορίες δεδομένων και μεταδεδομένων:

- Ποιότητα σχεδίασης και διαχείρισης: το πρώτο αναφέρεται στην ικανότητα ενός μοντέλου να αναπαριστά πληροφορίες με επάρκεια και αποτελεσματικά, ενώ το δεύτερο αναφέρεται στον τρόπο με τον οποίο το μοντέλο εξελίσσεται κατά τη λειτουργία της αποθήκης δεδομένων.
- Ποιότητα υλοποίησης λογισμικού: λαμβάνονται υπόψη οι διαστάσεις ποιότητας του προτύπου ISO 9126, καθώς η υλοποίηση λογισμικού δεν είναι εργασία με συγκεκριμένα χαρακτηριστικά αποθήκης δεδομένων.

- Ποιότητα χρήσης δεδομένων: αναφέρεται στις διαστάσεις που χαρακτηρίζουν τη χρήση και την αναζήτηση δεδομένων που περιέχονται στην αποθήκη δεδομένων.

Για κάθε διάσταση που περιέχεται στις άνωθεν κατηγορίες, εφαρμόζονται κατάλληλες μέθοδοι μέτρησης. Οι μέθοδοι αυτοί μαζί με τον βαθμό συνάφειας που σχετίζεται με κάθε διάσταση από τα ενδιαφερόμενα μέρη είναι τα δεδομένα για το αποτελεσματική μέτρηση. Στη φάση της αξιολόγησης ποιότητας, υπάρχει αποθήκευση των ακόλουθων πληροφοριών σχετικά με κάθε διάσταση ποιότητας δεδομένων:

1. απαιτήσεις ποιότητας (ένα διάστημα αναμενόμενων τιμών),
2. η επιτυγχανόμενη ποιοτική μέτρηση,
3. το μετρικό σύστημα που χρησιμοποιείται για τον υπολογισμό μιας μέτρησης,
4. πιθανές αιτιώδεις εξαρτήσεις από άλλες διαστάσεις ποιότητας. Οι πληροφορίες σχετικά με τις εξαρτήσεις μεταξύ των διαστάσεων ποιότητας χρησιμοποιούνται για τον εντοπισμό και την ανάλυση προβλημάτων ποιότητας. Ο προσδιορισμός των κρίσιμων περιοχών είναι το τελευταίο βήμα που αναλύεται στη μέθοδο. Στην πραγαμτικότητα, απλά αναφέρεται στη φάση βελτίωσης, αλλά δεν παρέχει εποικοδομητικές γνώσεις σχετικά με τον τρόπο βελτίωσης της ποιότητας μιας αποθήκης δεδομένων.

Ενότητα 4.3.3 – Μέθοδος Total Information Quality Management (TIQM)

Σε συνέχεια της παραπάνω μεθόδου, και η μέθοδος Total Information Quality Management (TIQM) [42] έχει σχεδιαστεί για να υποστηρίζει έργα αποθήκης δεδομένων. Η μεθοδολογία προϋποθέτει την ενοποίηση των λειτουργικών πηγών δεδομένων σε μια ολοκληρωμένη βάση δεδομένων, που χρησιμοποιείται σε όλους τους τύπους συναθροίσεων που εκτελούνται για την κατασκευή της αποθήκης δεδομένων. Αυτή η ενοποίηση εξαλείφει τα σφάλματα και τις ετερογένειες των βάσεων δεδομένων από τις οποίες προήλθαν τα δεδομένα. Η μέθοδος TIQM εστιάζει στις δραστηριότητες διαχείρισης που είναι υπεύθυνες για την ενοποίηση των λειτουργικών πηγών δεδομένων, συζητώντας τη στρατηγική που πρέπει να ακολουθήσει ένας οργανισμός προκειμένου να κάνει αποτελεσματικές τεχνικές επιλογές. Οι αναλύσεις κόστους-οφέλους υποστηρίζονται από μια διαχειριστική άποψη. Σημειώνεται ότι η μέθοδος TIQM παρέχει λεπτομερή ταξινόμηση ωφέλειας – κόστους .

Στην Εικόνα 4.3 παρουσιάζονται οι φάσεις της μεθόδου TIQM. Από διαχειριστικής άποψης στη μέθοδο, υπάρχουν τρεις φάσεις: Αξιολόγηση (Assessment), Βελτίωση (Improvement) και Διαχείριση & Παρακολούθηση βελτίωσης (Improvement Management & Monitoring). Μία από τις πολύτιμες συνεισφορές της μεθοδολογίας είναι ο ορισμός αυτής της τελευταίας φάσης, η οποία παρέχει κατευθυντήριες οδηγίες για τη διαχείριση των αλλαγών στη δομή του οργανισμού, σύμφωνα με τις απαιτήσεις διαχείρισης ποιότητας δεδομένων. Μάλιστα, η οικονομική προσέγγιση εισάγει την αξιολόγηση ωφέλειας – κόστους για να δικαιολογήσει τις παρεμβάσεις στην ποιότητα των δεδομένων. Στόχος δεν είναι μόνο η επίτευξη υψηλότερου επιπέδου ποιότητας δεδομένων, αλλά η ανάληψη βελτιωτικών ενεργειών μόνον εάν είναι εφικτές, δηλαδή μόνον εάν τα οφέλη είναι μεγαλύτερα από το κόστος.

Εικόνα 4.3 Φάσεις της μεθόδου TIQM.
(Πηγή: [36]).

Ενότητα 4.3.4 – Μέθοδος AIMQ (A Methodology for Information Quality Assessment)

Η μεθοδος AIMQ είναι η μόνη μεθοδολογία ποιότητας πληροφοριών που εστιάζει στη συγκριτική αξιολόγηση/προτυποποίηση (benchmarking) [43], προτείνοντας μια αντικειμενική και ανεξάρτητη από τον εκάστοτε τομέα τεχνική για την αξιολόγηση της ποιότητας δεδομένων.

Η μέθοδος AIMQ βασίζεται σε έναν πίνακα 2x2, που ονομάζεται προσομοίωμα PSP/IQ (Πίνακας 4.1), το οποίο ταξινομεί τις διαστάσεις ποιότητας, ανάλογα με τη σημασία τους από την οπτική γωνία του χρήστη και του διαχειριστή. Οι άξονες του πίνακα είναι η συμμόρφωση με τις προδιαγραφές και τις προσδοκίες των χρηστών. Ως εκ τούτου, διακρίνονται τέσσερις κατηγορίες διαστάσεων: ήχος (sound information), αξιόπιστη (dependable information), χρήσιμη (useful information) και χρησιμοποιήσιμη (usable

information) και οι διαστάσεις ποιότητας που προσδιορίζονται στους [44] ταξινομούνται σε αυτές τις κατηγορίες. Η συγκριτική αξιολόγηση πρέπει να κατατάσσει τις πληροφορίες σε κάθε μία από αυτές τις κατηγορίες.

Πίνακας 4.1 Προσομοίωμα PSP/IQ.
(Πηγή: [36]).

	Συμμόρφωση με προδιαγραφές	**Ικανοποίηση ή υπέρβαση προσδοκιών καταναλωτή**
Ποιότητα προϊόντος	Ηχητικές πληροφορίες	Χρήσιμες πληροφορίες
Ποιότητα υπηρεσίας	Αξιόπιστες πληροφορίες	Χρησιμοποιήσιμες πληροφορίες

Το προσομοίωμα PSP/IQ αποτελεί την είσοδο (input) στη μέθοδο AIMQ, της οποίας οι φάσεις συνοψίζονται στην Εικόνα 4.4. Η βιβλιογραφία που περιγράφει τη μέθοδο επικεντρώνονται κυρίως στις δραστηριότητες αξιολόγησης, ενώ δεν παρέχονται οδηγίες, τεχνικές και εργαλεία για δραστηριότητες βελτίωσης.

Εικόνα 4.4 Φάσεις της μεθόδου AIMQ.
(Πηγή: [36]).

Το IQ καταρτίζεται κυρίως μέσω ερωτηματολογίων. Ένα πρώτο πιλοτικό ερωτηματολόγιο χρησιμοποιείται για τον προσδιορισμό των σχετικών διαστάσεων ποιότητας και ιδιοτήτων που πρέπει να συγκριθούν. Στη συνέχεια, ένα δεύτερο ερωτηματολόγιο εξετάζει τις διαστάσεις και τα χαρακτηριστικά που είχαν προηγουμένως προσδιοριστεί προκειμένου να ληφθούν IQ μέτρα. Τέλος, τα μέτρα αυτά συγκρίνονται με δείκτες αναφοράς. Οι Lee et al. [2002] παρέχουν μια λίστα με τυπικές διαστάσεις και χαρακτηριστικά ποιότητας που βοηθούν στον ορισμό των ερωτηματολογίων.

Η βιβλιογραφία για τη μέθοδο AIMQ δεν παρέχει καμία περιγραφή της βάσης δεδομένων συγκριτικής αξιολόγησης που απαιτείται για την εφαρμογή της μεθοδολογίας. Οι τεχνικές ανάλυσης χάσματος (gap analysis) υποστηρίζονται ως τυπική προσέγγιση για τη διεξαγωγή συγκριτικής αξιολόγησης και την ερμηνεία των αποτελεσμάτων. Συγκεκριμένα, προτείνονται δύο τεχνικές ανάλυσης χάσματος:

- Ποιότητα πληροφορίας με Benchmark gaps, όπου γίνεται σύγκριση τις τιμές ποιότητας ενός οργανισμού με εκείνες των οργανισμών βέλτιστης πρακτικής.
- Ποιότητα πληροφορίας με Role gaps, όπου γίνεται σύγκριση των αξιολογήσεων ποιότητας των πληροφοριών που παρέχονται από διαφορετικούς οργανωτικούς ρόλους, δηλαδή τον επαγγελματία του πληροφοριακού συστήματος (IS professional) και τον χρήστη των πληροφοριών.

Τα IQ Role Gaps αναζητούν αποκλίσεις μεταξύ των αξιολογήσεων που παρέχονται από διαφορετικούς ρόλους ως ενδείξεις πιθανών ζητημάτων ποιότητας. Οι αποκλίσεις συνδέονται με μια κατεύθυνση. Η κατεύθυνση του χάσματος είναι θετική, εάν η αξιολόγηση των επαγγελματιών του πληροφοριακού συστήματος είναι υψηλότερη από την αξιολόγηση των χρηστών. Όμως, ένα μεγάλο θετικό χάσμα θεωρείται επικίνδυνο, καθώς δείχνει ότι οι επαγγελματίες του πληροφοριακού συστήματος δεν γνωρίζουν τα ζητήματα ποιότητας που έχουν εντοπίσει οι χρήστες πληροφοριών. Εάν το μέγεθος του χάσματος είναι μικρό, θα πρέπει να αναλυθεί η θέση του κενού. Εάν η τοποθεσία είναι υψηλή, υποδηλώνοντας υψηλό IQ, οι σταδιακές βελτιώσεις είναι πιο κατάλληλες, ενώ εάν η τοποθεσία είναι χαμηλή, είναι πιθανό να απαιτηθούν σημαντικές προσπάθειες βελτίωσης.

Ενότητα 4.3.5 – Μέθοδος DQA (Data Quality Assessment)

Στη βιβλιογραφία, οι μετρήσεις ποιότητας δεδομένων ορίζονται τις περισσότερες φορές κατά περίπτωση για την επίλυση συγκεκριμένων προβλημάτων και ως εκ τούτου εξαρτώνται από το εξεταζόμενο σενάριο. Η μέθοδος DQA [45] στοχεύει στον εντοπισμό των γενικών αρχών μέτρησης ποιότητας που είναι κοινές σε προηγούμενες έρευνες.

Η ταξινόμηση των μετρήσεων της μεθοδολογίας DQA συνοψίζεται στην Εικόνα 4.5. Η μεθοδολογία διακρίνει τις μετρήσεις ποιότητας, σε υποκειμενικές και αντικειμενικές. Οι υποκειμενικές μετρήσεις μετρούν τις αντιλήψεις, τις ανάγκες και τις εμπειρίες των ενδιαφερομένων. Στη συνέχεια, οι αντικειμενικές μετρήσεις ταξινομούνται σε ανεξάρτητες από εργασία και σε εξαρτημένες από εργασία. Οι πρώτες αξιολογούν την ποιότητα των δεδομένων χωρίς γνώση του πλαισίου της εφαρμογής. Οι τελευταίες ορίζονται για συγκεκριμένα πλαίσια εφαρμογής και περιλαμβάνουν επιχειρηματικούς κανόνες, εταιρικούς και κυβερνητικούς κανονισμούς και περιορισμούς που παρέχονται

από τη διαχείριση της βάσης δεδομένων. Και οι δύο μετρήσεις χωρίζονται σε τρεις κατηγορίες: απλή αναλογία, ελάχιστη ή μέγιστη τιμή και σταθμισμένος μέσος όρος.

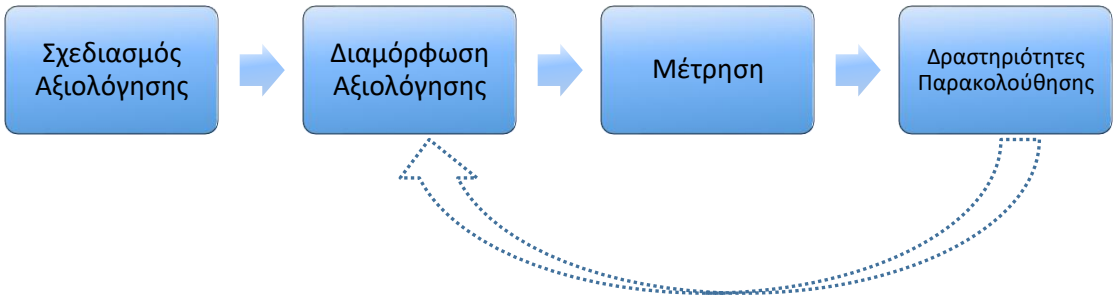
Εικόνα 4.5 Φάσεις της μεθόδου DQA.
(Πηγή: [36]).

Ενότητα 4.3.6 – Μέθοδος IQM (Information Quality Measurement)

Η μέθοδος IQM [46] στοχεύει στην παροχή ενός πλαισίου ποιότητας πληροφοριών προσαρμοσμένου στα δεδομένα του διαδικτύου. Συγκεκριμένα, η μέθοδος βοηθάει στην επιλογή και εξατομίκευση των εργαλείων που υποστηρίζουν τους προγραμματιστές στη δημιουργία, διαχείριση και συντήρηση τοποθεσιών του διαδικτύου.

Η μέθοδος IQM προσφέρει οδηγίες για να διασφαλίσει ότι τα εργαλεία λογισμικού αξιολογούν όλες τις θεμελιώδεις διαστάσεις της ποιότητας των πληροφοριών. Δύο πακέτα οδηγιών αναφέρονται: το πλαίσιο ποιότητας πληροφοριών που ορίζει τα κριτήρια ποιότητας και το σχέδιο δράσης που υποδεικνύει τον τρόπο εκτέλεσης μετρήσεων ποιότητας.

Οι κύριες φάσεις της μεθοδολογίας IQM αναφέρονται στην Εικόνα 4.6. Η πρώτη φάση ορίζει το σχέδιο μέτρησης. Το πλαίσιο ποιότητας πληροφοριών ορίζεται ως ένας κατάλογος κριτηρίων ποιότητας των πληροφοριών που προσδιορίζονται με συνεντεύξεις με τα ενδιαφερόμενα μέρη. Το πλαίσιο αυτό αποτελεί την είσοδο για έναν έλεγχο ποιότητας πληροφοριών, ο οποίος συσχετίζει τα κριτήρια ποιότητας των πληροφοριών με τις μεθόδους και τα εργαλεία που θα χρησιμοποιηθούν στη διαδικασία μέτρησης. Ορισμένα κριτήρια απαιτούν πολλαπλές μεθόδους μέτρησης. Η μεθοδολογία IQM συντονίζει την εφαρμογή πολλαπλών μεθόδων μέτρησης.



Εικόνα 4.6 Φάσεις της μεθόδου IQM.
(Πηγή: [36]).

Ενότητα 4.3.7 – Μέθοδος ISTAT (Italian National Bureau of Census)

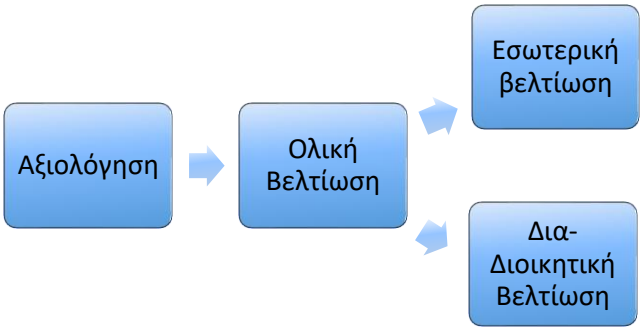
Η μέθοδος ISTAT [47, 48] σχεδιάστηκε στο πλαίσιο του Ιταλικού Εθνικού Γραφείου Απογραφής για τη συλλογή και διατήρηση υψηλής ποιότητας στατιστικών δεδομένων για Ιταλούς πολίτες και επιχειρήσεις. Το θεμελιώδες ζήτημα που προσπαθεί να διαχειριστεί η μέθοδος είναι πώς να εγγυηθεί την ποιότητα των δεδομένων που ενσωματώνονται από πολλαπλές βάσεις δεδομένων διαφορετικών τοπικών δημόσιων διοικήσεων. Το ζήτημα αυτό είναι ιδιαίτερα δύσκολο στο ιταλικό πλαίσιο, όπου η Δημόσια Διοίκηση είναι οργανωμένη σε τρία επίπεδα (Κεντρικό, Τοπικό, Περιφερειακό) και το καθένα διαχειρίζεται τα δικά του δεδομένα αυτόνομα. Η μεθοδολογία ISTAT εστιάζει στους πιο συνηθισμένους τύπους δεδομένων που ανταλλάσσονται μεταξύ διαφορετικών επιπέδων της Δημόσιας Διοίκησης, δηλαδή στα ιδιωτικά δεδομένα. Η μεθοδολογία επικεντρώνεται έντονα σε επίσημους κανόνες, καθώς στοχεύει στη ρύθμιση των δραστηριοτήτων διαχείρισης δεδομένων με τέτοιο τρόπο ώστε η ενσωμάτωσή τους να ικανοποιεί βασικές απαιτήσεις ποιότητας.

Οι κύριες φάσεις της μεθοδολογίας είναι οι εξής (Εικόνα 4.7):

- Η φάση αξιολόγησης (Assessment phase), η οποία εκτελείται αρχικά στις κεντρικές βάσεις δεδομένων που ανήκουν και διαχειρίζονται από την ISTAT, για τον εντοπισμό ζητημάτων ποιότητας.
- Η φάση της καθολικής βελτίωσης (Global Improvement phase), η οποία επιτελεί τη σύνδεση αρχείων μεταξύ των εθνικών βάσεων δεδομένων και το

σχεδιασμό της βελτιωτικής λύσης σε διαδικασίες, συμπεριλαμβανομένης της απόφασης λήψης, αγοράς ή προσαρμογής υπαρχουσών λύσεων.

- Δραστηριότητες βελτίωσης σε βάσεις δεδομένων που ανήκουν και διαχειρίζονται οι τοπικές διοικήσεις. Αυτές οι δραστηριότητες θα πρέπει να εκτελούνται από τις ίδιες τις τοπικές διοικήσεις, με τη βοήθεια εργαλείων και μαθημάτων που παρέχονται από την ISTAT.
- Δραστηριότητες βελτίωσης που απαιτούν τη συνεργασία πολλαπλών διοικήσεων. Αυτές οι δραστηριότητες είναι συνήθως προσανατολισμένες στη διαδικασία, καθώς αντιμετωπίζουν ροές δεδομένων που ανταλλάσσονται κατά την εκτέλεση συγκεκριμένων λειτουργικών δραστηριοτήτων. Ενδέχεται να απαιτούνται κεντρικές βάσεις δεδομένων για τον καλύτερο δυνατό συντονισμό. Οι δραστηριότητες αυτές σχεδιάζονται και συντονίζονται κεντρικά.



Εικόνα 4.7 Φάσεις της μεθόδου ISTAT.
(Πηγή: [36]).

Η μεθοδολογία ISTAT παρέχει ποικιλία στατιστικών τεχνικών για τη μέτρηση της ποιότητας. Παρέχει επίσης εργαλεία για τις πιο σχετικές δραστηριότητες καθαρισμού δεδομένων. Στη μεθοδολογία ISTAT, οι κάτοχοι δεδομένων ορίζονται με υψηλό επίπεδο λεπτομέρειας, που αντιστοιχεί σε ατομικά χαρακτηριστικά, όπως το MunicipalityCode. Η μεθοδολογία υποστηρίζει την τυποποίηση των μορφών δεδομένων και την έκφρασή τους σε ένα κοινό σχήμα XML, που επιτρέπει την ενοποίηση των βάσεων δεδομένων των τοπικών Διοικήσεων. Τα δεδομένα που ανταλλάσσονται μεταξύ διαφορετικών διοικήσεων επανασχεδιάζονται χρησιμοποιώντας μια αρχιτεκτονική λογισμικού που βασίζεται σε συμβάντα (στηρίζεται σε μηχανισμούς δημοσίευσης και εγγραφής).

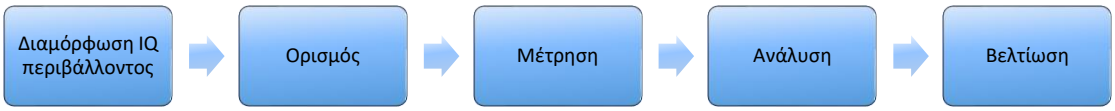
Ενότητα 4.3.8 – Μέθοδος AMEQ (Activity-based Measuring and Evaluating of Product information Quality)

Η μέθοδος AMEQ [49] προσφέρει μια αυστηρή βάση για την αξιολόγηση και τη βελτίωση της Ποιότητας Πληροφοριών Προϊόντος (Product Information Quality / PIQ), σε συμμόρφωση πάντα με τους στόχους του οργανισμού. Η μεθοδολογία ειδικεύεται στην αξιολόγηση της ποιότητας των δεδομένων σε εταιρείες του παραγωγικού κλάδου, όπου οι πληροφορίες προϊόντων αντιπροσωπεύουν το κύριο συστατικό της λειτουργίας των βάσεων δεδομένων. Επιπρόσθετα, στον εν λόγω κλάδο, η συσχέτιση μεταξύ πληροφοριών προϊόντος και διαδικασιών παραγωγής είναι απλή και σχετικά τυπική σε όλες τις εταιρείες, ενώ το σχήμα των βάσεων δεδομένων προϊόντων είναι επίσης παρόμοιο σε διαφορετικούς οργανισμούς.

Η μέθοδος αποτελείται από πέντε φάσεις για τη μέτρηση και τη βελτίωση του PIQ (Εικόνα 4.8):

- Η πρώτη φάση αξιολογεί την πολιτισμική ετοιμότητα ενός οργανισμού, χρησιμοποιώντας το Πλέγμα Ωριμότητας Διαχείρισης Ποιότητας Πληροφοριών, ένα πρότυπο για τη διεξαγωγή συνεντεύξεων για βασικούς διευθυντικούς ρόλους. Σε αυτή τη φάση, οι διαστάσεις του PIQ ορίζονται και ταξινομούνται ανάλογα με τη συνάφειά τους για διαφορετικές επιχειρηματικές δραστηριότητες.
- Η δεύτερη φάση προσδιορίζει το προϊόν πληροφορίας. Κάθε προϊόν πληροφοριών συνδέεται με μια αντίστοιχη επιχειρηματική διαδικασία, που μοντελοποιείται μέσω μιας αντικειμενοστρεφούς προσέγγισης. Στη μεθοδολογία AMEQ, μοντελοποιούνται οκτώ τύποι αντικειμένων: ανθρώπινοι πόροι, πόροι πληροφοριών, επιχειρηματικές δραστηριότητες, εισροές πόρων, διεργασίες πόρων, εκροές πόρων, μέτρα απόδοσης και επιχειρηματικοί στόχοι. Σε αυτή τη φάση, παράγεται επίσης μοντέλο μεθόδων μέτρησης.
- Η τρίτη φάση εστιάζει στις μετρήσεις.
- Η τέταρτη φάση, διερευνά αιτίες για πιθανά ζητήματα του PIQ, αναλύοντας τις διαστάσεις ποιότητας που έχουν λάβει χαμηλή βαθμολογία.
- Η πέμπτη φάση είναι η φάση βελτίωσης του PIQ.

Για την τέταρτη και την πέμπτη φάση, η μέθοδος AMEQ δεν προσφέρει μεθόδους και εργαλεία λειτουργίας, αλλά μόνο γενικές οδηγίες.



Εικόνα 4.8 Φάσεις της μεθόδου AMEQ.
(Πηγή: [36]).

Ενότητα 4.3.9 – Μέθοδος COLDQ (Cost-Effect Of Low Data Quality)

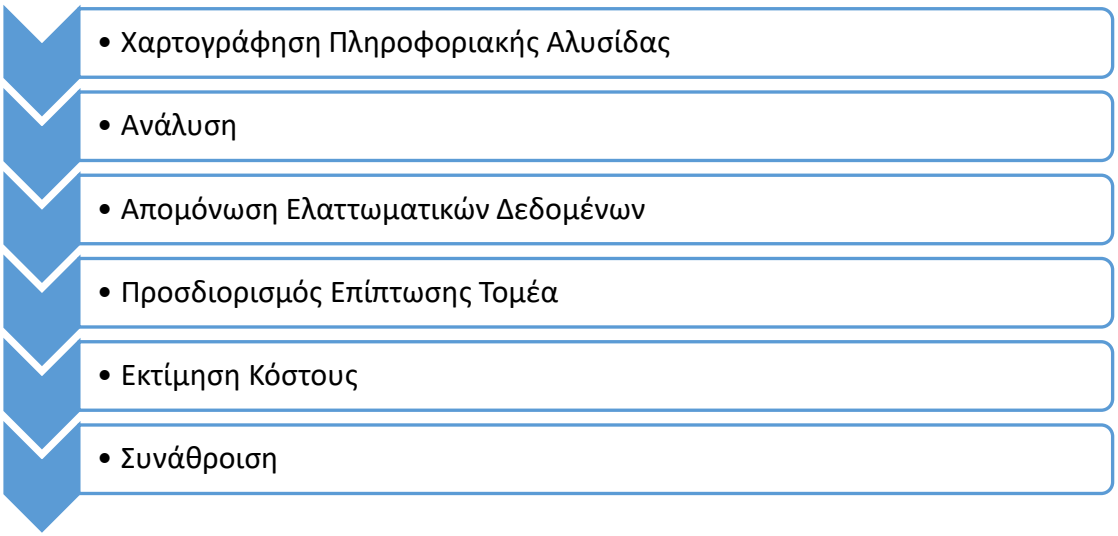
Στόχος της μεθόδου COLDQ [50] είναι να παρέχει μια κάρτα βαθμολογίας ποιότητας δεδομένων, λαμβάνοντας υπόψη την επίδραση του κόστους της χαμηλής ποιότητας δεδομένων. Άμεσα οφέλη επιτυγχάνονται από την αποφυγή δαπανών από κακής ποιότητας δεδομένα, μέσω της υιοθέτησης τεχνικών βελτίωσης. Σκοπός είναι να επιτευχθεί μια ποσοτική αξιολόγηση του βαθμού στον οποίο οι επιχειρηματικές διαδικασίες επηρεάζονται από κακές πληροφορίες.

Η μέθοδος αποτελείται από έξι φάσεις για την αξιολόγηση του κόστους-αποτελέσματος της χαμηλής ποιότητας δεδομένων (Εικόνα 4.9). Στην πρώτη φάση της μεθόδου, το επιχειρηματικό πλαίσιο προσομοιώνεται προσδιορίζοντας δύο μοντέλα ροής δεδομένων:

- τη στρατηγική ροή δεδομένων, που χρησιμοποιείται για τη λήψη αποφάσεων,
- και τη λειτουργική ροή δεδομένων, που χρησιμοποιείται για την επεξεργασία δεδομένων.

Και τα δύο μοντέλα αντιπροσωπεύουν ένα σύνολο σταδίων επεξεργασίας που περιγράφουν τη ροή πληροφοριών από την παροχή δεδομένων στην κατανάλωση δεδομένων. Με βάση αυτά τα μοντέλα, διεξάγονται οι αντικειμενικές και υποκειμενικές αναλύσεις του επιχειρηματικού πλαισίου. Εσωτερικοί και εξωτερικοί χρήστες, υπάλληλοι και πελάτες πραγματοποιούν συνεντεύξεις προκειμένου να εντοπιστούν ελαττωματικά δεδομένα. Στη συνέχεια, υπολογίζονται τα σφάλματα σε ελαττωματικές δραστηριότητες στα στρατηγικά και επειχειρησιακά μοντέλα του

επιχειρηματικού πλαισίου. Αυτή η συσχέτιση μεταξύ σφαλμάτων και δραστηριοτήτων παρέχει τη βάση για αξιολογήσεις κόστους. Η μεθοδολογία COLDQ παρέχει μια ταξινόμηση των λειτουργικών, τακτικών και στρατηγικών οικονομικών επιπτώσεων που πρέπει να ληφθούν υπόψη. Σε κάθε κατηγορία κόστους αποδίδεται μια οικονομική αξία με βάση τη γνώση του πλαισίου. Το κόστος αντιπροσωπεύει την εισροή στην τελική φάση βελτίωσης. Η μεθοδος COLDQ υποστηρίζει αναλύσεις ωφέλειας – κόστους αξιολογώντας και συγκεντρώνοντας το κόστος των έργων βελτίωσης της ποιότητας. Μάλιστα, οι κατευθυντήριες γραμμές που δίδονται, υποστηρίζουν τον υπολογισμό της απόδοσης της εκάστοτε επένδυσης (ROI) και του νεκρού σημείου των βελτιωτικών ενεργειών.



Εικόνα 4.9 Φάσεις της μεθόδου COLDQ.
(Πηγή: [36]).

Ενότητα 4.3.10 – Μέθοδος DaQuinCIS (Data Quality In Cooperative Information Systems)

Η μέθοδος DaQuinCIS [51] αντιμετωπίζει ζητήματα ποιότητας δεδομένων στα Συνεργατικά Πληροφοριακά Συστήματα (Cooperative Information Systems). Η συνεργασία εγείρει δύο ζητήματα ποιότητας δεδομένων που σχετίζονται με το συγκεκριμένο πλαίσιο. Πρώτον, η ποιότητα των δεδομένων βασίζεται στην εμπιστοσύνη μεταξύ των οργανισμών. Δεύτερον, η κακή ποιότητα δεδομένων μπορεί να εμποδίσει τη συνεργασία και, ως εκ τούτου, έχει εκτεταμένες συνέπειες. Για την αντιμετώπιση του πρώτου ζητήματος, η μεθοδολογία DaQuinCIS εισάγει την έννοια

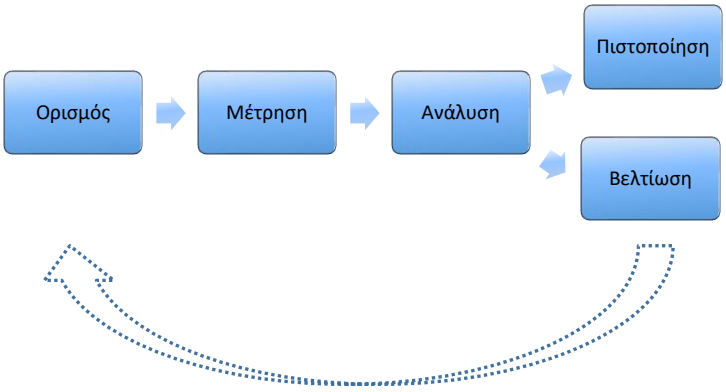
της πιστοποίησης ποιότητας δεδομένων, η οποία συσχετίζει τα δεδομένα με αντίστοιχα μέτρα ποιότητας που ανταλλάσσονται μεταξύ των οργανισμών μαζί με τα δεδομένα. Το δεύτερο ζήτημα αντιμετωπίζεται με την παροχή μηχανισμών επιλογής δεδομένων με βάση την ποιότητα. Αυτοί οι μηχανισμοί επιλογής προσδιορίζουν τα δεδομένα υψηλότερης ποιότητας μεταξύ επικαλυπτόμενων βάσεων δεδομένων που ανήκουν σε διαφορετικούς συνεργαζόμενους οργανισμούς. Με αυτόν τον τρόπο, η συνεργασία βοηθάει στη βελτίωση της ποιότητας.

Η μεθοδολογία DaQuinCIS παρέχει ένα καινοτόμο μοντέλο για την αναπαράσταση της ποιότητας δεδομένων που ονομάζεται Δεδομένα και Ποιότητα Δεδομένων (D^2Q), το οποίο περιλαμβάνει:

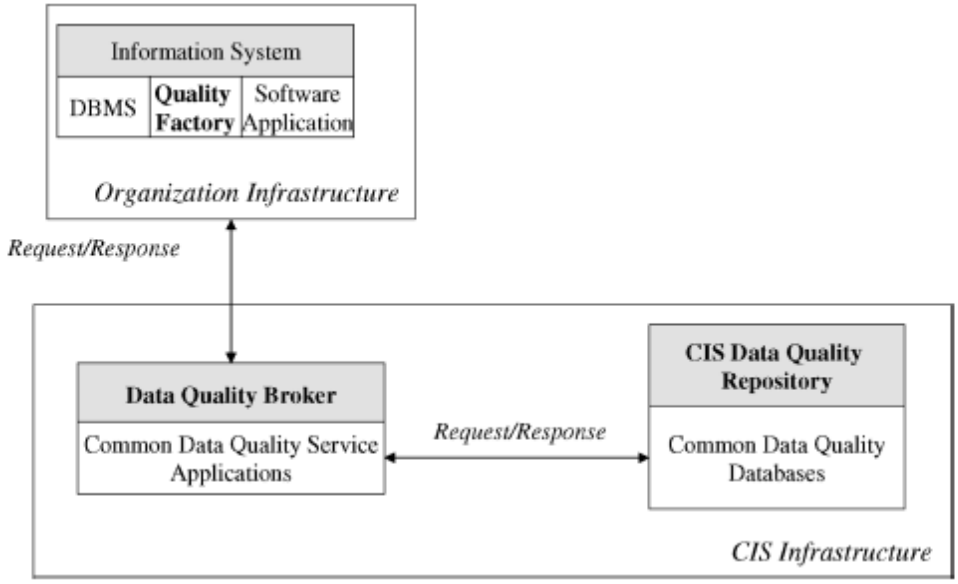
- δομές για την αναπαράσταση δεδομένων,
- ένα σύνολο ιδιοτήτων ποιότητας δεδομένων,
- δομές για την αναπαράσταση ιδιοτήτων ποιότητας δεδομένων,
- συσχετισμούς μεταξύ δεδομένων και ποιοτικών μεταδεδομένων.

Η αξιοπιστία της πηγής περιλαμβάνεται μεταξύ των ποιοτικών ιδιοτήτων. Η τιμή που σχετίζεται με αυτήν την ιδιότητα, εκχωρείται από έναν τρίτο εξωτερικό οργανισμό με βάση διάφορες παραμέτρους, συμπεριλαμβανομένου του αριθμού των παραπόνων που υποβλήθηκαν από άλλους οργανισμούς και του αριθμού των αιτημάτων που εκδόθηκαν σε κάθε πηγή.

Η Εικόνα 4.10 παρουσιάζει τις θεμελιώδεις φάσεις της μεθοδολογίας DaQuinCIS: ανάλυση ποιότητας, αξιολόγηση ποιότητας, πιστοποίηση ποιότητας και βελτίωση ποιότητας. Η μεθοδολογία υποστηρίζεται από μια αρχιτεκτονική που αποτελείται από μια εσωτερική και μια εξωτερική υποδομή (βλ. Εικόνα 4.11). Οι φάσεις υλοποιούνται από αντίστοιχες ενότητες του «Εργοστασίου» Ποιότητας (Quality Factory / CF) που υλοποιούνται σε κάθε οργανισμό που ανήκει στο CIS. Η επικοινωνία μεταξύ των οργανισμών οι οποίοι συμμετέχουν στο CIS ενεργοποιείται από μια υπηρεσία ποιοτικής ειδοποίησης που είναι μια μηχανή δημοσίευσης/συνδρομής, που χρησιμοποιείται ως γενικός δίαυλος μηνυμάτων μεταξύ των διαφορετικών αρχιτεκτονικών στοιχείων.



Εικόνα 4.10 Φάσεις της μεθόδου DaQuinCIS.
(Πηγή: [36]).



Εικόνα 4.11 Η «αρχιτεκτονική» της μεθόδου DaQuinCIS.
(Πηγή: [36]).

Στο CF, τα αιτήματα από εξωτερικούς χρήστες ορίζουν δεδομένα προς ανάκτηση και τις αντίστοιχες απαιτήσεις ποιότητας. Μια κατάλληλη ενότητα αξιολογεί την ποιότητα των δεδομένων και συγκρίνει τις τιμές ποιότητας με τις απαιτήσεις ποιότητας που εκφράζονται από τους χρήστες. Εάν τα δεδομένα δεν ικανοποιούν τις απαιτήσεις ποιότητας, αποστέλλεται ειδοποίηση στον χρήστη. Αντίθετα, εάν οι τιμές ποιότητας είναι ικανοποιητικές, ένα πιστοποιητικό ποιότητας συσχετίζεται με τα δεδομένα και αποστέλλεται στον χρήστη. Η βελτίωση της ποιότητας πραγματοποιείται από τον Διακομιστή Ποιότητας Δεδομένων (Data Quality Broker). Αυτή η ενότητα, σε συνεργασία με το Αποθετήριο Ποιότητας Δεδομένων (Data Quality Repository),

μεταφράζει ερωτήματα σύμφωνα με ένα παγκόσμιο σχήμα και επιλέγει δεδομένα που μεγιστοποιούν την ποιότητα. Ένα ερώτημα που υποβάλλεται από έναν συγκεκριμένο οργανισμό εκδίδεται σε όλους τους οργανισμούς, προσδιορίζοντας ένα σύνολο απαιτήσεων ποιότητας για τα ζητούμενα δεδομένα. Διαφορετικά αντίγραφα των ίδιων δεδομένων που λαμβάνονται ως απάντηση στο αίτημα, εναρμονίζονται και ως εκ τούτου, επιλέγονται οι τιμές καλύτερης ποιότητας οι οποίες και επιστρέφονται στους αιτούντες οργανισμούς.

Όσον αφορά τις τεχνικές που χρησιμοποιούνται στη φάση βελτίωσης, η μέθοδος προτείνει έναν νέο αλγόριθμο για την αντιστοίχιση αρχείων. Το Record Matcher είναι ένα στοιχείο του Data Quality Broker. Το Record Matcher εφαρμόζει μια μέθοδο αντιστοίχισης εγγραφών, με βάση τα δεδομένα ποιότητας που εξάγονται από συνεργαζόμενους οργανισμούς.

Ενότητα 4.3.11 – Μέθοδος QAFD (Quality Assessment of Financial Data)

Η μεθοδολογία QAFD [52] ορίζει πρότυπα ποιοτικά μέτρα για τα οικονομικά επιχειρησιακά δεδομένα και έτσι ελαχιστοποιεί το κόστος των εργαλείων μέτρησης ποιότητας. Η μεθοδολογία συνδυάζει ποσοτικές αντικειμενικές και ποιοτικές υποκειμενικές αξιολογήσεις για τον εντοπισμό ζητημάτων ποιότητας και την επιλογή των κατάλληλων ενεργειών βελτίωσης της ποιότητας. Καθορίζονται δείκτες που εξαρτώνται από το περιβάλλον, κανόνες ποιότητας δεδομένων, μετρήσεις και στρατηγικές για ποσοτικές και ποιοτικές αξιολογήσεις.

Οι κύριες φάσεις αυτής της μεθοδολογίας αναφέρονται στην Εικόνα 4.12. Πρώτον, η μεθοδολογία επιλέγει τις πιο σχετικές χρηματοοικονομικές μεταβλητές. Η επιλογή βασίζεται συνήθως σε γνώσεις από προηγούμενες αξιολογήσεις, ανάλογα με την πρακτική αποτελεσματικότητά τους. Οι μεταβλητές ομαδοποιούνται σε κατηγορίες που επηρεάζουν όμοια τη συμπεριφορά των επενδυτών και των καταναλωτών και χαρακτηρίζονται από τους ίδιους επιχειρηματικούς και περιγραφικούς παράγοντες, καθώς και παράγοντες κινδύνου.

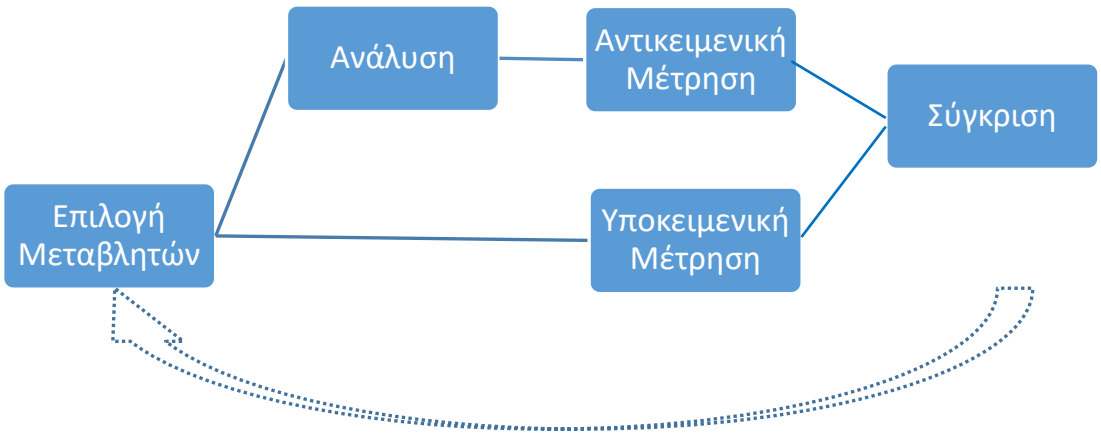
Η δεύτερη φάση στοχεύει στην ανακάλυψη των βασικών αιτιών των σφαλμάτων. Σε αυτή τη φάση προσδιορίζονται οι πιο σχετικές διαστάσεις ποιότητας δεδομένων και

παράγονται κανόνες ποιότητας δεδομένων. Οι κανόνες ποιότητας δεδομένων αντιπροσωπεύουν τις δυναμικές σημασιολογικές ιδιότητες των μεταβλητών που δεν μπορούν να μετρηθούν, σύμφωνα με τις διαστάσεις ποιότητας.

Στην τρίτη φάση, η αντικειμενική αξιολόγηση πραγματοποιείται με βάση ποσοτικούς δείκτες. Οι συγγραφείς προτείνουν ένα μαθηματικό μοντέλο για την αντικειμενική αξιολόγηση με αποτέλεσμα μια συνολική κατάταξη των δεδομένων σε κάθε διάσταση ποιότητας.

Η υποκειμενική αξιολόγηση πραγματοποιείται στην τέταρτη φάση από τρεις διαφορετικές οπτικές γωνίες: εμπειρογνώμονες επιχειρήσεων, πελάτες και ειδικούς στην ποιότητα δεδομένων. Κάθε ερωτώμενος πρέπει να αξιολογήσει το επίπεδο ποιότητας σε κάθε διάσταση ποιότητας. Η συνολική αξιολόγηση λαμβάνεται ως η μέση τιμή της υποκειμενικής αξιολόγησης κάθε κατηγορίας εμπειρογνωμόνων.

Στην πέμπτη και τελευταία φάση συγκρίνονται αντικειμενικές και υποκειμενικές αξιολογήσεις. Για κάθε διάσταση υπολογίζεται η διαφορά μεταξύ της αντικειμενικής και της υποκειμενικής αξιολόγησης. Εάν η διαφορά είναι θετική, η αντικειμενική αξιολόγηση πρέπει να επανεξεταστεί για να επισημανθούν ζητήματα ποιότητας που σχετίζονται με την άποψη των ειδικών.



Εικόνα 4.12 Φάσεις της μεθόδου QAFD.
(Πηγή: [36]).

Ενότητα 4.3.12 – Μέθοδος CIHI (Canadian Institute for Health Information)

Η μέθοδος CIHI είναι μια μέθοδος για την αξιολόγηση και τη βελτίωση της ποιότητας των δεδομένων του Καναδικού Ινστιτούτου Πληροφοριών Υγείας [53]. Κατά τη μέθοδο, το κύριο ζήτημα είναι το μέγεθος των βάσεων δεδομένων και η ετερογένειά τους. Η μεθοδολογία CIHI υποστηρίζει την επιλογή ενός υποσυνόλου δεδομένων για την εστίαση της φάσης αξιολόγησης ποιότητας, ενώ προτείνει ένα ευρύ σύνολο κριτηρίων ποιότητας για την αξιολόγηση της ετερογένειας.

Η στρατηγική για την ποιότητα δεδομένων CIHI προτείνει μια προσέγγιση δύο φάσεων (Εικόνα 4.13). Η πρώτη φάση είναι ο ορισμός ενός Πλαισίου Ποιότητας Δεδομένων (Data Quality Framework / DQF) και η δεύτερη είναι μια εις βάθος ανάλυση των δεδομένων με τη μεγαλύτερη επισκεψιμότητα. Το DQF ορίζεται σε τρία βήματα:

1) Τυποποίηση των πληροφοριών ποιότητας δεδομένων.
2) Ανάπτυξη κοινής στρατηγικής για την αξιολόγηση της ποιότητας των δεδομένων.
3) Ορισμός διαδικασιών εργασίας για τη διαχείριση δεδομένων, οι οποίες προσδιορίζουν προτεραιότητες ποιότητας δεδομένων και εφαρμόζουν διαδικασίες συνεχούς βελτίωσης δεδομένων.

Η εφαρμογή της μεθοδού CIHI είναι κυκλική, σύμφωνα με μια προσέγγιση της συνεχούς βελτίωσης. Για την επιτυχή υλοποίηση απαιτούνται τα ακόλουθα:

- Ορισμός της χρονικής περιόδου για έναν κύκλο.
- Καθορισμός χρονικών στόχων για διαφορετικούς ποιοτικούς στόχους.
- Κατανομή ad hoc πόρων για ανάλυση, αξιολόγηση και τεκμηρίωση ποιότητας δεδομένων.
- Κατανομή ad hoc πόρων για τη διαδικασία βελτίωσης της ποιότητας των δεδομένων.

Η ανάλυση των δεδομένων που χρησιμοποιούνται πιο συχνά πραγματοποιείται σε τρία βήματα: Ανάλυση ποιότητας δεδομένων, Μέτρηση και Τεκμηρίωση/Αναφορά (Εικόνα 4.13). Τα έγγραφα της Τεκμηρίωσης/Αναφοράς αναφέρουν τα προβλήματα ποιότητας που εντοπίστηκαν από την ανάλυση και την Μέτρηση της ποιότητας των δεδομένων.

Εικόνα 4.13 Φάσεις ανάλυσης τακτικών δεδομένων της μεθόδου CIHI.
(Πηγή: [36]).

Η αξιολόγηση της ποιότητας των δεδομένων βασίζεται σε ένα ιεραρχικό μοντέλο τεσσάρων επιπέδων. Σε πρώτο επίπεδο ορίζονται 86 βασικά κριτήρια ποιότητας. Αυτά τα κριτήρια συγκεντρώνονται μέσω αλγορίθμων σύνθεσης σε 24 ποιοτικά χαρακτηριστικά στο δεύτερο ιεραρχικό επίπεδο και περαιτέρω συγκεντρώνονται σε πέντε διαστάσεις ποιότητας στο τρίτο επίπεδο. Τέλος, οι πέντε διαστάσεις συγκεντρώνονται σε μία συνολική αξιολόγηση βάσης δεδομένων στο τέταρτο επίπεδο.

Σημειώνεται, ότι η αρχική και βασική αξιολόγηση των 86 κριτηρίων ποιότητας δεδομένων πραγματοποιείται μέσω ερωτηματολογίων που αναφέρουν κριτήρια ως στοιχεία που θα βαθμολογηθούν σε μια τακτική κλίμακα τεσσάρων βαθμών ως «μη εφαρμόσιμο», «άγνωστο», «δεν πληρούται» ή «πληρούται». Στη συνέχεια, σε κάθε επίπεδο συγκέντρωσης, οι αξιολογήσεις ελέγχονται. Η διαδικασία επικύρωσης διασφαλίζει ότι η ερμηνεία και η βαθμολόγηση κάθε κριτηρίου είναι όσο το δυνατόν πιο τυπική.

Ενότητα 4.3.13 – Μέθοδος CDQ (Complete Data Quality)

Η μεθοδολογία CDQ [54,55] έχει σχεδιαστεί με γνώμονα την πληρότητα, την ευελιξία και την απλότητα στην εφαρμογή. Η πληρότητα επιτυγχάνεται με την εξέταση των υπαρχουσών τεχνικών και εργαλείων και την ενσωμάτωσή τους σε ένα πλαίσιο που μπορεί να λειτουργήσει τόσο σε ενδο-επιχειρησιακά όσο και σε δια-επιχειρησιακά πλαίσια και μπορεί να εφαρμοστεί σε όλους τους τύπους δεδομένων (δομημένων, ημιδομημένων και μη). Η μεθοδολογία είναι ευέλικτη, αφού υποστηρίζει τον χρήστη στην επιλογή των καταλληλότερων τεχνικών και εργαλείων σε κάθε φάση και σε οποιοδήποτε πλαίσιο. Επίσης, η μέθοδος CDQ υποστηρίζει την απλότητα, αφού

οργανώνεται σε φάσεις και κάθε φάση χαρακτηρίζεται από συγκεκριμένο στόχο και σύνολο τεχνικών προς εφαρμογή.

Η μέθοδος CDQ παρέχει υποστήριξη για την επιλογή της βέλτιστης διαδικασίας βελτίωσης ποιότητας που μεγιστοποιεί τα οφέλη εντός δεδομένων ορίων προϋπολογισμού. Δίνει έμφαση στη φάση δημιουργίας αρχικών απαιτήσεων. Η εστίαση είναι στο πώς να επιτευχθεί η συνολική ποιότητα δεδομένων χωρίς να παρέχονται ενδείξεις σχετικά με τον τρόπο χρήσης της πλαισιακής γνώσης. Ένας από τους στόχους της μεθόδου είναι να παρουσιαστεί μια ποσοτική αξιολόγηση του βαθμού στον οποίο οι επιχειρηματικές διαδικασίες επηρεάζονται από κακές πληροφορίες.

Πιο αναλυτικά, τρεις κύριες φάσεις χαρακτηρίζουν τη μεθοδολογία: ανασυγκρότηση κατάστασης, αξιολόγηση και επιλογή της βέλτιστης διαδικασίας διόρθωσης (Εικόνα 4.14).

Στην πρώτη φάση της μεθοδολογίας, ανακατασκευάζονται οι σχέσεις μεταξύ οργανωτικών μονάδων, διαδικασιών, υπηρεσιών και δεδομένων. Αυτές οι σχέσεις μοντελοποιούνται χρησιμοποιώντας πίνακες που περιγράφουν ποιες οργανωτικές μονάδες χρησιμοποιούν δεδομένα και τους ρόλους τους στις διάφορες επιχειρηματικές διαδικασίες. Επιπλέον, σε αυτή τη φάση περιγράφονται οι διαδικασίες μαζί με τη συμβολή τους στην παραγωγή αγαθών/υπηρεσιών και τους νομικούς και οργανωτικούς κανόνες που ορίζουν τις ροές εργασίας.

Η δεύτερη φάση θέτει νέα επίπεδα στόχων ποιότητας που απαιτούνται για τη βελτίωση της ποιότητας των διαδικασιών και αξιολογεί τα αντίστοιχα κόστη και οφέλη. Αυτή η φάση εντοπίζει τις κρίσιμες μεταβλητές που επηρεάζονται από την κακή ποιότητα. Δεδομένου, ότι οι δραστηριότητες βελτίωσης είναι περίπλοκες και δαπανηρές, η μέθοδος εστιάζει στα μέρη των βάσεων δεδομένων και στις ροές δεδομένων που δημιουργούν τα μεγαλύτερα ζητήματα.

Η τρίτη φάση αποτελείται από πέντε βήματα και στοχεύει στον προσδιορισμό της βέλτιστης διαδικασίας βελτίωσης ή αλλιώς της σειράς δραστηριοτήτων που έχει την υψηλότερη αναλογία κόστους/αποτελεσματικότητας. Τα νέα επίπεδα ποιότητας στόχου ορίζονται λαμβάνοντας υπόψη το κόστος και τα οφέλη. Μπορούν να πραγματοποιηθούν διάφορες δραστηριότητες βελτίωσης για την επίτευξη νέων ποιοτικών στόχων. Η μεθοδολογία συνιστά τον προσδιορισμό όλων των τεχνικών

βελτίωσης που βασίζονται στα δεδομένα και στις διαδικασίες (data-driven ή process-driven) για τις βάσεις δεδομένων που επηρεάζονται από την κακή ποιότητα. Ένα σύνολο αμοιβαία συνεπών τεχνικών βελτίωσης συνιστά μια διαδικασία βελτίωσης. Τέλος, επιλέγεται η καταλληλότερη διαδικασία βελτίωσης με ανάλυση ωφέλειας – κόστους.

Εικόνα 4.14 Φάσεις της μεθόδου CDQ.
(Πηγή: [36]).

Ενότητα 4.3.14 – Μέθοδος HDQM (Heterogeneous Data Quality Methodology)

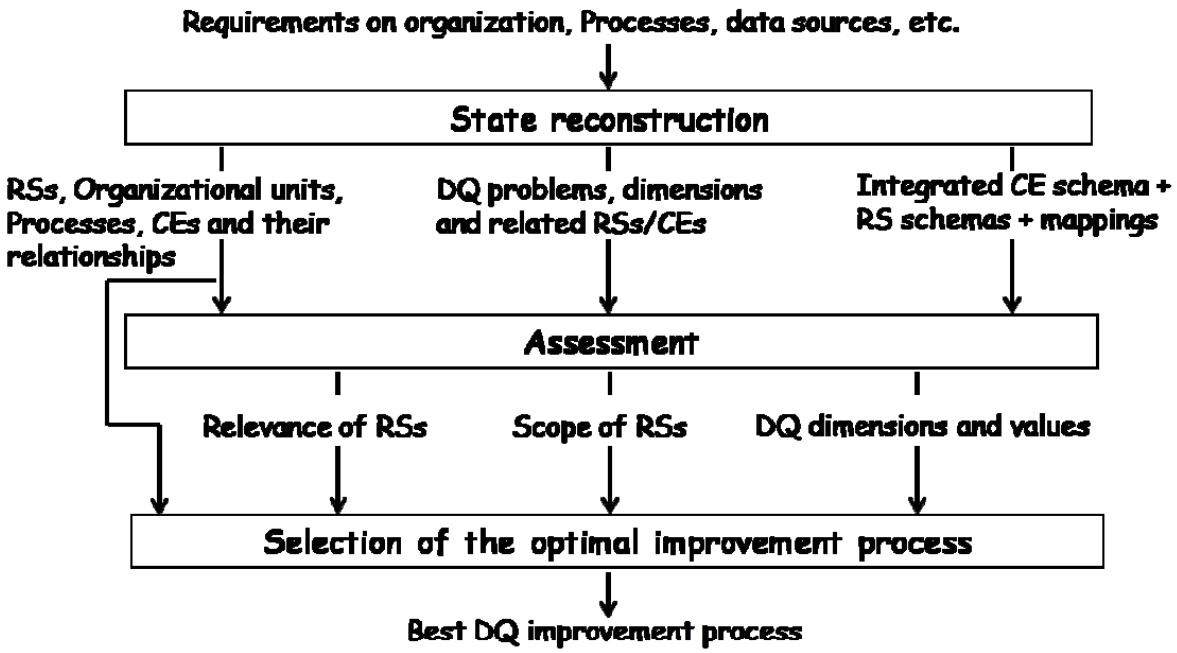
Οι Batini et al. [56] παρουσίασαν μια Μεθοδολογία Ποιότητας Ετερογενών Δεδομένων (HDQM) για την αξιολόγηση και τη βελτίωση της ποιότητας δεδομένων που λαμβάνει υπόψη όλους τους πιθανούς τύπους δεδομένων που διαχειρίζεται ένας οργανισμός, όπως δομημένα δεδομένα (βάσεις δεδομένων), ημιδομημένα δεδομένα (συνήθως σε γλώσσα XML) και μη δομημένα δεδομένα (έγγραφα). Ορίζεται επίσης ένα μετα-μοντέλο για να περιγραφεί η σχετική γνώση που διαχειρίζεται η μέθοδος. Συνοπτικά, η HDQM παρουσιάζει ένα σύνολο φάσεων για την αξιολόγηση και τη βελτίωση της ποιότητας των δεδομένων.

Πιο συγκεκριμένα, η μέθοδος HDQM θεωρείται επέκταση της μεθόδου CDQ σε ημιδομημένες και μη δομημένες πηγές δεδομένων. Οι μη επεξεργασμένοι πίνακες σε έγγραφα, λίστες στοιχείων ή αρχεία, όπου συγκεντρωτικά σύνολα τιμών διαχωρίζονται με κάποιο οριοθέτη (π.χ. κόμμα ή άνω και κάτω τελεία) είναι συχνά το πρώτο στάδιο στο οποίο τα σχεσιακά δεδομένα διαχειρίζονται από άτομα, πριν από την κωδικοποίησή τους σε αρχεία XML (για ηλεκτρονική ανταλλαγή) ή βάσεις δεδομένων (για αναζήτηση και αποθήκευση). Στα μη δομημένα σχεσιακά δεδομένα, δεν δίνονται πληροφορίες σχετικά με τους τύπους δεδομένων και τους περιορισμούς, εκτός από το απλό περιεχόμενο της ίδιας της σχέσης (ή του πίνακα). Για το λόγο αυτό, διαφέρουν από τα ημιδομημένα έγγραφα, καθώς δε συνδέονται με κανένα σαφώς καθορισμένο σχήμα. Η μέθοδος εστιάζει σε σχεσιακά δεδομένα επιτρέποντας την υπόθεση ότι οι

πόροι πληροφοριών που εξετάζονται ορίζονται τόσο με εντατικούς όρους, δηλαδή καλύπτουν και καθιστούν σαφείς διαφορετικές πτυχές των οργανωτικών εννοιών, όσο και σε επεκτατικούς όρους, δηλαδή συσχετίζουν διαφορετικές αξίες με διαφορετικές πτυχές της έννοιας που αντιπροσωπεύουν. Σημειώνεται ότι το εύρος της έρευνας δεν καλύπτει ζητήματα που αντιμετωπίζονται από το ευρύ φάσμα επιστημονικών κλάδων που διερευνούν μη δομημένα δεδομένα κειμένου, όπως η επεξεργασία φυσικής γλώσσας και η ανάκτηση πληροφοριών, ή άλλες μη δομημένες πηγές (π.χ. εικόνα και ήχος).

Η μέθοδος HDQM έχει ως κύριο στόχο να προσφέρει ενδείξεις για το καλύτερο πρόγραμμα βελτίωσης ποιότητας δεδομένων που πρέπει να αναλάβει ένας οργανισμός σε σχέση με τις ιδιαίτερες ανάγκες και περιορισμούς του. Η μέθοδος αποτελείται από τρεις κύριες φάσεις και καθεμία από αυτές αποτελείται από έναν αριθμό βημάτων (Εικόνα 4.15). Ειδικότερα, οι κύριες φάσεις είναι:

1. Ανασυγκρότηση της κατάστασης (State reconstruction), που στοχεύει στην ανασυγκρότηση όλης της σχετικής γνώσης σχετικά με τις οργανωτικές μονάδες, τις διαδικασίες, τους πόρους και τις εννοιολογικές οντότητες που εμπλέκονται στον οργανισμό.
2. Αξιολόγηση, η οποία στοχεύει στην απόκτηση ποσοτικής αξιολόγησης προβλημάτων ποιότητας δεδομένων. Οι διαστάσεις ποιότητας δεδομένων μετρώνται για να αξιολογηθεί το τρέχον επίπεδο ποιότητας δεδομένων και να οριστούν οι νέοι στόχοι ποιότητας που πρέπει να επιτευχθούν στο τέλος του προγράμματος βελτίωσης.
3. Βελτίωση, όπου οι κατάλληλες δραστηριότητες βελτίωσης επιλέγονται αξιολογώντας τις επιπτώσεις τους ως προς την αναλογία Διάσταση ποιότητας δεδομένων/Κόστος.



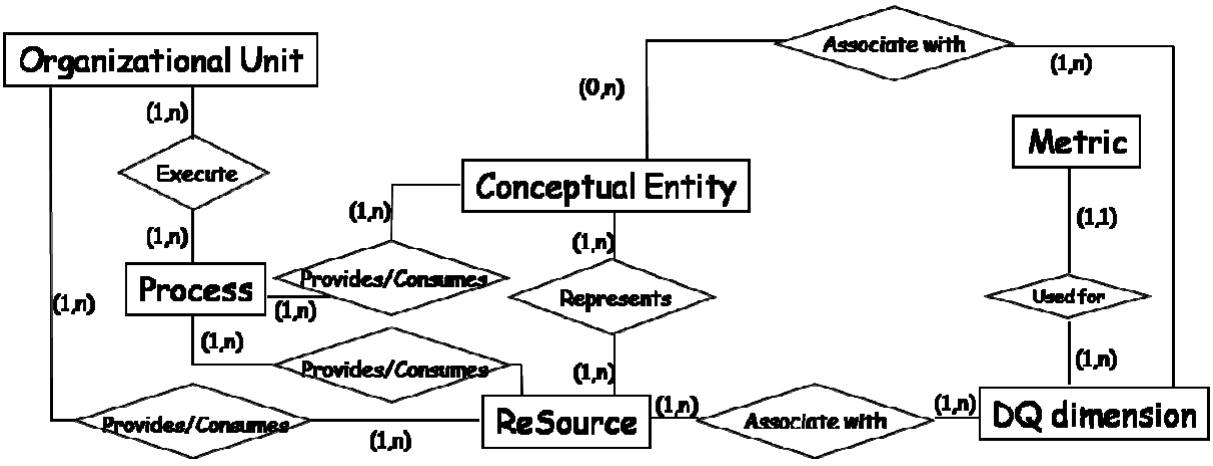
Εικόνα 4.15 Φάσεις της μεθόδου HDQM, πόροι εισόδου και πόροι εξόδου.
(Πηγή: [56]).

Όπως αναφέρθηκε παραπάνω, η HDQM παρουσιάζει ένα μετα-μοντέλο (Εικόνα 4.16). Συνοπτικά, παρουσιάζονται τα εξής στοιχεία:

- Η Διαδικασία (Process) είναι μια αρθρωμένη ακολουθία δραστηριοτήτων που εκτελούνται από και εντός του εξεταζόμενου οργανισμού. Οι διαδικασίες λαμβάνονται υπόψη, καθώς η αποτελεσματικότητά τους επηρεάζεται από την ποιότητα των δεδομένων.
- Η Οργανωτική Μονάδα (Organizational Unit) είναι ένα σημαντικό στοιχείο ενός οργανισμού που εμπλέκεται στην παραγωγή, χρήση και επεξεργασία δεδομένων. Χαρακτηρίζεται από μια συγκεκριμένη εσωτερική δομή και ένα σύνολο εσωτερικών κανόνων.
- Ο πόρος (ReSource/RS) είναι οιαδήποτε πηγή πληροφοριών που ένας οργανισμός μπορεί είτε να χρησιμοποιήσει, είτε να αποκτήσει πρόσβαση σε αυτή. Οι πόροι μπορούν να θεωρηθούν, είτε επιχειρησιακά περιουσιακά στοιχεία, είτε η ίδια η προέλευση αυτών των δεδομένων (δηλαδή η πηγή). Μέσα σε έναν επιχειρηματικό οργανισμό, τα RS διαχειρίζονται κατά τη διάρκεια διαδικασιών, είτε παραγωγής πληροφοριών, είτε κατανάλωσης πληροφοριών. Τυπικά παραδείγματα RS είναι βάσεις δεδομένων, ροές δεδομένων,

ηλεκτρονικά ή έντυπα έγγραφα. Τα δεδομένα που αντιπροσωπεύονται από τον όρο RS μπορεί να είναι δομημένα, ημιδομημένα ή μη δομημένα.

- Η Εννοιολογική οντότητα (Conceptual Entity / CE) είναι κάθε έννοια που αναφέρεται σε ένα μεμονωμένο φαινόμενο της πραγματικότητας ενδιαφέροντος και που είναι δυνατόν να αφαιρεθεί από τα RS που χρησιμοποιούνται σε έναν οργανισμό: π.χ. πελάτης, προμηθευτής, εγκατάσταση, εμπόρευμα. Το CE αναφέρεται σε μια έννοια που είναι ανεξάρτητη από τον συγκεκριμένο τρόπο που την αντιπροσωπεύει ένα μεμονωμένο RS, καθώς και από το φυσικό μέσο και τη μορφή του RS.
- Το μετα-μοντέλο της μεθόδου HDQM προτείνει μια ρητή διάκριση μεταξύ RS και CE, επειδή πρέπει να εξεταστούν και οι διαστάσεις που σχετίζονται με την ποιότητα των RS και την ποιότητα των CE, καθεμία με τις πιο κατάλληλες τεχνικές για την υπό εξέταση διάσταση. Επιπλέον, ο στενός δεσμός μεταξύ των RS και των αντίστοιχων CE (που αντιπροσωπεύονται σε αυτό) πρέπει να ληφθεί υπόψη για να επιτρέψει και να υποστηρίξει μια ολοκληρωμένη προσέγγιση για την αξιολόγηση και τη βελτίωση της ποιότητας των δεδομένων. Στο μετα-μοντέλο της μεθόδου, ένα CE μπορεί να αναφέρεται από πολλά RS μέσα σε έναν οργανισμό και αντίστροφα.



Εικόνα 4.16 Μετα-μοντέλο μεθόδοου HDQM.
(Πηγή: [56]).

Επιπρόσθετα για την μέτρηση της ποιότητας των δεδομένων χρησιμοποιούνται δύο διαστάσεις: Ακρίβεια (Accuracy) και Μοντερνισμός (Currency). Για τον εντοπισμό

σφαλμάτων και τη διόρθωση των δεδομένων χρησιμοποιούνται συγκεκριμένοι αλγόριθμοι βάσει στρατηγικής βασιζόμενης στα δεδομένα.

Εν κατακλείδι, η μέθοδος HDQM:

- Παρέχει κατευθυντήριες γραμμές για την εξέταση διαφορετικών τύπων δεδομένων, για την ανάλυση της ποιότητας των πληροφοριών που διαχειρίζεται ένας οργανισμός και την επιλογή αποτελεσματικών λύσεων για τη βελτίωση της ποιότητας των δεδομένων. Ωστόσο, η εύρεση επαρκούς αντιστάθμισης μεταξύ απλότητας και πληρότητας είναι ένα δύσκολο έργο που εξαρτάται σε μεγάλο βαθμό από τον τομέα της εφαρμογής και τη συγκεκριμένη περίπτωση.
- Είναι μια μεθοδολογία υψηλού επιπέδου και γενικής εφαρμογής που προτείνεται από τους δημιουργούς περισσότερο ως πλαίσιο γνωστών τεχνικών και μεθόδων παρά ως μια νέα εργαλειοθήκη.
- Οι μετρήσεις ποσοτικής αξιολόγησης απαιτούν σημαντική προσπάθεια, ενώ απαιτούν μια προσέγγιση πολλαπλών δυαδικών πλευρών και δεν αξιοποιούν μια στατιστική (βασισμένη σε δείγμα) προσέγγιση. Από την άλλη πλευρά, οι ποιοτικές μετρήσεις βασίζονται στη διαθεσιμότητα και κυρίως στην ικανότητα των ειδικών του τομέα.
- Η φάση της Ανασυγκρότησης, όπως έχει συλληφθεί στο πλαίσιο της μεθόδου, είναι μια χρονοβόρα και απαιτητική δραστηριότητα πόρων. Αυτό ισχύει ιδιαίτερα για την εξαγωγή του σχήματος των σχετικών πόρων, εάν δεν είναι ακόμη καλά γνωστοί και τεκμηριωμένοι. Ωστόσο, είναι εξαιρετικά ευέλικτη στην εξέταση μεμονωμένων πόρων και μπορεί να εφαρμοστεί αποτελεσματικά σε τμήματα του συνόλου των πόρων που χρησιμοποιούνται στον οργανισμό.

Ενότητα 4.3.15 – Ανάλυση των μεθόδων VORD & DAQUAVORD

Η αναγνώριση της ποιότητας ενός λογισμικού εξαρτάται από μεταβλητές, όπως η ποιότητα των πληροφοριών και των δεδομένων που διαχειρίζεται η εφαρμογή [57-62]. Αυτό αναγνωρίζεται από τα πρότυπα SQUARE (ISO/IEC 25000), τα οποία υπογραμμίζουν την ανάγκη αντιμετώπισης της ποιότητας των δεδομένων ως μέρος της αξιολόγησης του επιπέδου ποιότητας του προϊόντος λογισμικού, σύμφωνα με τα οποία «το σύστημα υπολογιστή-στόχου περιλαμβάνει επίσης υλικό υπολογιστή, μη

στοχευμένα προϊόντα λογισμικού, μη στοχευμένα δεδομένα και τα δεδομένα στόχου, τα οποία αποτελούν το αντικείμενο εξέτασης του μοντέλου ποιότητας δεδομένων» [63].

Βάσει του παραπάνω, οι οργανισμοί θα ήταν καλό να λαμβάνουν υπόψη τις ανησυχίες για την ποιότητα των δεδομένων κατά την ανάπτυξη διαφόρων λογισμικών, καθώς τα δεδομένα αποτελούν βασικό παράγοντα [64-66]. Επομένως, τέτοιες ανησυχίες για την ποιότητα των δεδομένων θα μπορούσαν να ληφθούν υπόψη στο αρχικό στάδιο ανάπτυξης, όπως θα ίσχυε για κάθε άλλη απαίτηση. Συμπεριλαμβανομένου του [67, 68], δεν υπάρχουν σημαντικές προτάσεις που να επικεντρώνονται στην αντιμετώπιση προβλημάτων ποιότητας δεδομένων στη διαδικασία ανάπτυξης λογισμικού.

Εισάγεται, λοιπόν, η έννοια της Απαίτησης Λογισμικού Ποιότητας Δεδομένων (Data Quality Software Requirement / DQSR), ως μια μέθοδος για την εφαρμογή μιας Απαίτησης Ποιότητας Δεδομένων (Data Quality Requirement / DQR) σε μια εφαρμογή [68]. Το DQSR περιγράφεται ως απαίτηση λογισμικού που στοχεύει στην ικανοποίηση ενός DQR. Η αιτιολόγηση της ιδέας είναι η εξής: για την καταγραφή των DQR που ταιριάζουν καλύτερα με τα δεδομένα που χρησιμοποιεί ένας χρήστης σε κάθε σενάριο χρήσης και αργότερα, να δημιουργηθούν τα DQSR που θα συμπληρώσουν τις κανονικές απαιτήσεις λογισμικού που συνδέονται με καθένα από αυτά τα σενάρια. Η αντιμετώπιση πολλαπλών DQSR είναι μια πολύπλοκη διαδικασία, λαμβάνοντας υπόψη την ύπαρξη ισχυρών εξαρτήσεων, όπως εσωτερικοί περιορισμοί και αλληλεπίδραση με εξωτερικά συστήματα και την ποικιλομορφία των χρηστών. Ως αποτέλεσμα, τείνουν να επηρεάζουν και να δείχνουν τις συνέπειες των αντιφατικών επικαλύψεων, τόσο στα μοντέλα διεργασιών όσο και στα μοντέλα δεδομένων.

Όσον αφορά αυτή την πολυπλοκότητα και προσπαθώντας να βελτιωθούν οι προσπάθειες ανάπτυξης σχετικού λογισμικού, οι César Guerra-García et al. (2023) [114] εισήγαν τη μέθοδο DAQUAVORD, μια Μεθοδολογία για τη Διαχείριση Απαιτήσεων Ποιότητας Δεδομένων, η οποία βασίζεται στη μέθοδο Ορισμού Απαιτήσεων Προσανατολισμένης στην Απόψη (Viewpoint-Oriented Requirements Definition / VORD) και πιο πρόσφατο και αποδεκτό Πρότυπο ISO/IEC 25012. Η μέθοδος παρουσιάζεται ως καθολική και εύκολα προσαρμόσιμη σε διαφορετικά συστήματα πληροφοριών, όσον αφορά τόσο τη φύση τους, τον αριθμό και την ποικιλία των παραγόντων.

Αν και υπάρχουν πολλοί τρόποι για να ορίσουμε την ποιότητα των δεδομένων, στην προκειμένη περίπτωση, κάθε χρήστης, ο οποίος διαδραματίζει συγκεκριμένο ρόλο σε έναν οργανισμό που χρησιμοποιεί λογισμικό, θα πρέπει να μπορεί να προσδιορίσει έναν αριθμό διαστάσεων Ποιότητας Δεδομένων (Data Quality / DQ) που είναι σημαντικές για την αξιολόγηση του επιπέδου ποιότητας ενός συνόλου δεδομένων που πρέπει να χρησιμοποιήσει ο χρήστης για να εκτελέσει την εργασία. Ένα σύνολο διαστάσεων DQ που αναγνωρίζονται από τον χρήστη συνιστά Απαίτηση Ποιότητας Δεδομένων [68]. Οι αντίστοιχες Απαίτησης Ποιότητας Δεδομένων DQR θα πρέπει να καταγράφονται για κάθε σενάριο χρήσης και για κάθε χρήστη σε αυτά σενάρια.

Στη διεθνή βιβλιογραφία υπάρχουν μοντέλα ποιότητας δεδομένων [69] που εξαρτώνται σε μεγάλο βαθμό από τον τομέα αναφοράς των δεδομένων. Για παράδειγμα, υπάρχουν διαφορετικά μοντέλα ποιότητας δεδομένων σε τομείς όπως:

- Υγειονομική ή/και ιατρική περίθαλψη [70–74].
- Στρατιωτική αξιολόγηση και αξιολόγηση κινδύνου, συμπεριλαμβανομένων των τομέων που σχετίζονται με χημικούς κινδύνους [75, 76].
- Συστήματα υποστήριξης αποφάσεων [77, 78].
- Διαδικτυακές εφαρμογές [79, 80].
- Μικρές επιχειρήσεις [81].
- Συνεργατικά συστήματα [82].

Ταυτοχρόνως, υπάρχουν ορισμένα πολλαπλών χρήσεων μοντέλα Ποιότητας Δεδομένων, συμπεριλαμβανομένου του [82], που ασχολούνται με μια αντικειμενοστρεφής προσέγγιση, όπου επίκεντρο είναι ο χρήστης για την αξιολόγηση της ποιότητας των δεδομένων. Άλλα πιο παραδοσιακά παραδείγματα ποιότητας δεδομένων που βασίζονται σε διαστάσεις θεωρούνται πρότυπα, π.χ. Strong et al. [83], Redman [84], Dama [85].

Η υπόψη μεθοδολογία ορίζει την έννοια του DQSR ως «απαιτήσεις λογισμικού που μπορεί να προκύψουν από μια ενιαία απαίτηση ποιότητας δεδομένων (DQR) που επιτρέπει την εξαγωγή των απαραίτητων χαρακτηριστικών για το υπό ανάπτυξη λογισμικό και την υποστήριξη των απαιτούμενων χαρακτηριστικών ποιότητας δεδομένων». Έτσι, το χαρακτηριστικό DQ θα πρέπει να γίνει κατανοητό ως ένα σύνολο χαρακτηριστικών λογισμικού που υποστηρίζουν το συγκεκριμένο DQR.

Η ομαδοποίηση αυτών των χαρακτηριστικών DQ παρουσιάζεται με κοινό τρόπο από το πρότυπο ISO 25012 [86]. Για παράδειγμα, η ακρίβεια (accuracy), ερμηνεύεται σύμφωνα και με το ISO 25012, ως ένα σύνολο χαρακτηριστικών λογισμικού που στοχεύουν στην εγγύηση ότι τα δεδομένα που χρησιμοποιούνται αντιπροσωπεύουν την πραγματική τιμή του υπό εξέταση χαρακτηριστικού σε ένα συγκεκριμένο πλαίσιο.

Για να επιτευχθεί αυτό, τα DQSR:

- θα προκύψουν μερικές φορές σε νέα χαρακτηριστικά ή ακόμη και λειτουργίες που συμπληρώνουν την υπάρχουσα λειτουργική απαίτηση (π.χ. εάν ένας χρήστης συμπληρώνει μια φόρμα, θα πρέπει να προστεθεί μια συγκεκριμένη επαλήθευση ότι όλες οι τιμές εισάγονται για υποχρεωτικά πεδία),
- θα προέλθει από νέες μη λειτουργικές απαιτήσεις (π.χ. εάν επιλεγεί κατάλληλος τύπος δεδομένων για να διασφαλιστεί ότι οι αποθηκευμένες τιμές είναι αρκετά ακριβείς) ή,
- θα περιλαμβάνει και τα δύο είδη.

Η επιτυχία της ανάπτυξης λογισμικού εξαρτάται σε μεγάλο βαθμό από το πόσο καλά προσδιορίζονται οι απαιτήσεις λογισμικού από τους αναλυτές συστημάτων [87-91]. Είναι, επίσης, σημαντική η μοντελοποίηση τόσο των δεδομένων όσο και των απαιτήσεων που εμπλέκονται στη διαμόρφωση των Προδιαγραφών Απαιτήσεων Λογισμικού (Software Requirements Specification / SRS). Σημειώνεται ότι αρκετοί χρήστες θα χρησιμοποιούν το νέο λογισμικό και οι ίδιες λειτουργίες πρέπει να καλύπτουν όλες τις πιθανές ανάγκες των χρηστών, δηλαδή να πληρούνται όλες οι προβλεπόμενες απαιτήσεις. Για το σκοπό αυτό, η μέθοδος Ορισμού Απαιτήσεων Προσανατολισμένης στην Άποψη (Viewpoint-Oriented Requirements Definition/VORD) είναι καλά αναγνωρισμένη ως μέσο καθορισμού απαιτήσεων [92–95], ακόμη και αν αυτή η μέθοδος είναι προσανατολισμένη μόνο στη λειτουργικότητα. Έχει αποδειχθεί, επίσης, αποτελεσματική στον κλάδο της Μηχανικής για το Industry 4.0.

Στη μέθοδο VORD, η έννοια της οπτικής γωνίας/άποψης είναι παρόμοια με το ρόλο των χρηστών στην περίπτωση μιας εφαρμογής. Η εφαρμογή παρέχει λειτουργικότητα σε οπτικές γωνίες/απόψεις και αργότερα, οι οπτικές γωνίες/απόψεις μπορούν να εισφέρουν στην εφαρμογή πληροφορίες ελέγχου και σχετικές παραμέτρους [96]. Έτσι,

μια οπτική γωνία/άποψη θεωρείται ως μια εξωτερική οντότητα της εφαρμογής, η οποία μπορεί να δημιουργήσει μια απαίτηση.

Ένα από τα βασικά σημεία της μεθόδου VORD είναι ότι εξετάζει την ύπαρξη διαφορετικών οπτικών και προσφέρει ένα πλαίσιο για τη διαχείριση όλων των διαφορετικών απαιτήσεων, επιπλέον της ανακάλυψης πιθανών συγκρούσεων που θα μπορούσαν να προκύψουν μεταξύ προτεινόμενων απαιτήσεων από διαφορετικές απόψεις. Σύμφωνα με τη μέθοδο VORD που προτείνουν οι Kotonya και Sommerville [97], ακολουθούνται τα εξής βήματα:

1. Αναγνώριση απόψεων
2. Δόμηση οπτικής γωνίας
3. Τεκμηρίωση-εγγραφή απόψεων
4. Χαρτογράφηση συστήματος οπτικής γωνίας

Αντικείμενο της μεθόδου DAQUAVORD είναι να καθοδηγεί τους προγραμματιστές εφαρμογών, μέσω των δραστηριοτήτων εξαγωγής και προδιαγραφής DQR. Αυτή η δραστηριότητα περιλαμβάνει μια προδιαγραφή των δεδομένων που θα χρησιμοποιηθούν σε κάθε πιθανό σενάριο και τα αντίστοιχα σχετικά DQR που πρέπει να πληρούνται. Κάθε λειτουργία λογισμικού χρησιμοποιείται από έναν χρήστη που έχει ένα συγκεκριμένο ρόλο (οι λεγόμενες «Απόψεις», σύμφωνα με τη μέθοδο VORD). Υποδεικνύεται συγκεκριμένο DQR για τα δεδομένα που χρησιμοποιεί ο χρήστης. Αυτά τα DQR θα πρέπει να συλλέγονται, να ταξινομούνται, να ιεραρχούνται και να μετατρέπονται εύκολα σε DQSR, τα οποία θα πρέπει να εισαχθούν αργότερα ως τυπικές απαιτήσεις λογισμικού.

Συγκεκριμένα, η μέθοδος DAQUAVORD περιλαμβάνει έξι δραστηριότητες, λαμβάνοντας υπόψη τα βήματα που ορίζονται στη μέθοδο VORD, όπως ορίστηκαν παραπάνω. Αυτές είναι:

1. Σχεδιασμός του έργου για τις προδιαγραφές απαιτήσεων λογισμικού, οι οποίες αναπτύσσεται περαιτέρω σε πέντε υποδραστηριότητες.
2. Αναγνώριση οπτικής γωνίας/άποψης συστήματος πληροφοριών στη μέθοδο VORD. Αυτή η δραστηριότητα αναπτύσσεται περαιτέρω σε τρεις υποδραστηριότητες.
3. Δόμηση οπτικής γωνίας στη μέθοδο VORD. Αυτή η δραστηριότητα αναπτύσσεται περαιτέρω σε δύο υποδραστηριότητες.

4. Έγγραφη τεκμηρίωση οπτικών γωνιών στη μέθοδο VORD. Αυτή η δραστηριότητα αναπτύσσεται περαιτέρω σε δύο υποδραστηριότητες.
5. Καθορισμός της διάταξης των λειτουργιών του λογισμικού και των απαιτήσεων ποιότητας δεδομένων, ώστε να είναι δυνατή η χαρτογράφηση συστήματος των οπτικών γωνιών στη VORD. Αυτή η δραστηριότητα αναπτύσσεται περαιτέρω σε τρεις υποδραστηριότητες.
6. Ολοκλήρωση της προδιαγραφής των απαιτήσεων ποιότητας δεδομένων για το λογισμικό. Η δραστηριότητα αναπτύσσεται περαιτέρω σε δύο υποδραστηριότητες.

Ενότητα 4.4 – Συγκριτική ανάλυση μεθόδων και σχετικής βιβλιογραφίας

Όπως προκύπτει και από τις παραπάνω αναλύσεις, το ζήτημα της ανάλυσης και αξιολόγησης της ποιότητας των δεδομένων είναι πολυπαραγοντικό και εξαρτάται από τον τομέα στον οποίο αναφέρεται. Ως εκ τούτου, είναι λογικό διαφορετικές μέθοδοι και προσεγγίσεις ποιότητας δεδομένων να παρουσιάζονται ανά τομέα δεδομένων, ενώ κάποιες μέθοδοι παρουσιάζουν μια πιο ολιστική προσέγγιση. Για παράδειγμα, όπως προκύπτει και από τον Πίνακα 4.2, μέθοδοι που σύμφωνα με τους δημιουργούς τους, αφορούν την ανάλυση ποιότητας δεδομένων διαφόρων τομέων, είναι η TDQM, η AIMQ, η DQA, η ISTAT και άλλες, ενώ μέθοδος που ασχολείται με τα οικονομικά δεδομένα είναι η QAFD. Στον υπόψη Πίνακα 4.2, καταγράφονται (συνολικά 43) κλασικές μέθοδοι (γενικής ή ειδικής εφαρμογής), καθώς και αρκετά πρόσφατες βιβλιογραφικές αναφορές για μελέτες ποιότητας δεδομένων σε συγκεκριμένους τομείς:

- Αεροπορία (Aviation)
- Big Data
- Βιομηχανία κατασκευής (Construction Industry)
- Data Warehousing
- Deep learning
- Περιβάλλον (Environment)
- Οικονομία (Finance)
- Γεωλογία (Geology)

- Υγεία (Health management)
- Πληροφοριακά συστήματα (Information systems)
- Ανάπτυξη λογισμικού και διαδικτυακών εφαρμογών (Software / Web applications).

Ο αριθμός και ο τύπος των διαστάσεων που εφαρμόζονται για την εξέταση των δεδομένων ποικίλλει, ανάλογα τη μέθοδο. Στον Πίνακα 4.2 παρουσιάζονται οι διαστάσεις για κάθε μέθοδο. Είναι αντιληπτό, ότι οι πιο διαδεδομένες διαστάσεις ποιότητας είναι η ακρίβεια (accuracy) και η πληρότητα (completeness), μετρήσιμες με ποικίλους τρόπους μετρικών. Οι περισσότερες μέθοδοι και μελέτες, εστιάζουν σε λίγες κύριες διαστάσεις για τις αναλύσεις τους, έτσι ώστε να μειώνεται ο όγκος, ο χρόνος και η πολυπλοκότητα των υπολογισμών στο στάδιο της εξετάσης της ποιότητας των δεδομένων, με απότελεσμα να προκύπτει και πιο ξεκάθαρη αξιολόγηση των αποτελεσμάτων στο στάδιο της αξιολόγησης (assessment stage). Οι διαστάσεις σημειώνονται στον πίνακα με την λατινική ορολογία για λόγους συνέπειας με τη διατιθέμενη βιβλιογραφία.

Μεγαλύτερο ενδιαφέρον παρουσιάζει ο Πίνακας 4.3, όπου παρουσιάζεται, σε πρώτη φάση, ο τύπος τον δεδομένων που μπορεί να εξεταστεί από τις μεθόδους: Δομημένα (Structured/S), Ημι-δομημένα (Semi-structured/SM) και Αδόμητα (Unstructured/US). Είναι προφανές ότι το μεγαλύτερο ποσοστό των μεθόδων αφορά δομημένα ή/και ημι-δομημένα δεδομένα, καθώς:

- η διαχείριση τους απαιτεί λιγότερες ανάγκες προγραμματισμού,
- απαιτούνται λιγότερα μεταδεδομένα,
- απαιτούνται λιγότερες ανάγκες αποθήκευσης,
- πιθανόν αποδίδουν μικρότερα σφάλματα επεξεργασίας.

Χαρακτηριστικό παράδειγμα αποτελεί η μέθοδος HDQM, που αποτελεί την επέκταση της CDQ σε ημι-δομημένα και αδόμητα δεδομένα, προσφέροντας νέες διαδικασίες και τεχνικές ανάλυσης για τις απαιτήσεις της εν λόγω επέκτασης.

Στον Πίνακα 4.3 παρουσιάζεται ο τύπος της στρατηγικής ή των στρατηγικών (εάν αναφέρεται ξεκάθαρα) που εφαρμόζεται σε κάθε μέθοδο ή βιβλιογραφική αναφορά. Φαίνεται ότι στρατηγικές βασιζόμενες στα δεδομένα (data-driven) είναι πιο διαδεδομένες, καθώς παρέχουν ένα σύνολο αρκετών τεχνικών για την επεξεργασία των

δεδομένων, χωρίς αυτό να σημαίνει ότι οι δύο στρατηγικές δεν μπορούν να συνδυαστούν. Εξάλλου, μια μέθοδος που εφαρμόζει και τους δύο τύπους στρατηγικών μπορεί να εφαρμοστεί ευκολότερα σε έναν οργανισμό, καθώς ανεξάρτητα από τα δεδομένα, οι διαδικασίες καθορίζουν καίρια την ποιότητα των δεδομένων. Επίσης, πολλές φορές η αλλαγή μιας διαδικασίας μπορεί να επιφέρει θεαματικά αποτελέσματα στην ποιότητα των δεδομένων, από απλές διαδικασίες διόρθωσης πακέτων δεδομένων μιας λανθάνουσας πηγής.

Στην επόμενη στήλη του Πίνακα 4.3, παρουσιάζονται οι τεχνικές που εφαρμόζονται ανά μέθοδο ή μελέτη. Όπως διακρίνεται, παρουσιάζεται όλο το φάσμα των διαθέσιμων τεχνικών και για τα δύο είδη στρατηγικών. Δεσπόζουσα θέση κατέχουν οι τεχνικές της data-driven στρατηγικής, data normalization, data cleaning και record linkage. Ωστόσο, ανάλογα με τις απαιτήσεις της εκάστοτε εφαρμογής παρουσιάζεται χρήση και ιδιαίτερων και πιο σπάνιων τεχνικών όπως, η Synthetic Minority Over-sampling Technique (SMOTE) ή η sensor monitoring.

Στη συνέχεια παρουσιάζονται οι μέθοδοι ανάλυσης δεδομένων που αναφέρονται στις μελέτες, όπου παρουσιάζεται εκτεταμένη χρήση των μεθόδων μηχανικής μάθησης, όπως τα νευρωνικά δίκτυα, τα decision trees, η linear regression και η Support Vector Machines (SVM) καθώς και των στατιστικών μεθόδων ή μεθόδων clustering (k-means). Σημειώνεται ότι σε κάποιες μελέτες ακολουθούνται μεθολογίες ISO, ενώ πλέον παρουσιάζονται και μέθοδοι ανάλυσης τεχνητής νοημοσύνης (data-centric AI). Στην παρακάτω Εικόνα 4.17, παρουσιάζονται συνοπτικά τεχνικές και μέθοδοι ανάλυσης δεδομένων που παρουσιάζονται στις βιβλιογραφικές αναφορές.

Τεχνικές (Data-driven/Process-driven)	Data cleaning, Data transformation, Data approximation, Data filtering, Data Normalization, Record linkage, Error localization and correction, Data and schema integration, Cost optimization, Source trustworthiness, Source improvement, Backtranslation, Word embeddings, Outlier detection, Duplicate removal, Error correction, Synthetic Minority Over-sampling Technique (SMOTE), Data Augmentation, Sensor monitoring, Innovative signal processing technologies, Data interoperability, Data fusion, Data-centric AI, LOF-IDW cleaning method / Process control, Process redesign etc.
Μέθοδοι ανάλυσης δεδομένων	DNN, FCNN, CNN, LSTM, SVR, SVM, Bayesian Networks, Bayesian Optimization-SVM, LR, RF, GBDT, BLR, NN, PS-AE-LSTM, Data-centric AI, Clustering, k-means etc.

Εικόνα 4.17 Τεχνικές και μέθοδοι ανάλυσης βιβλιογραφικών αναφορών.

Στην τελευταία στήλη του πίνακα, παρουσιάζεται μια αναφορά στην ύπαρξη εκτίμησης (ή έστω αναφοράς) και αξιολόγηση του κόστους. Παρατηρείται ότι λίγες μελέτες τοποθετούν ως προτεραιότητα το επωμιζόμενο κόστος (υπολογιστικό κόστος, κόστος πόρων, χρηματικό κόστος, χρονικό κόστος κλπ.) από την εκτέλεση της αναλύσης ποιότητας δεδομένων ή/και της κακής ποιότητας των λανθασμένων δεδομένων που προκύπτουν. Η εν λόγω παράμετρος, αν και απαιτεί πρόσθετους χειρισμούς και πόρους, θα μπορούσε να βοηθήσει σημαντικά στην τροποποίηση των διαδικασιών και τη μείωση του τελικού κόστους της επιχείρησης ή του οργανισμού.

Πίνακας 4.2 Μέθοδοι και μελέτες ανάλυσης ποιότητας δεδομένων βιβλιογραφικής ανασκόπησης και διαστάσεις ποιότητας.

Α/Α	Μέθοδος / Μελέτη	Βιβλιογραφική αναφορά	Τομέας	Αριθμός διαστάσεων ποιότητας	Διαστάσεις ποιότητας
1	Total Data Quality Management (TDQM)	Wang 1998 [38]	General	15	Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of manipulation, Value added, Free of error, Interpretability, Objectivity, Relevance, Reputation, Security, Timeliness, Understandability
2	The Datawarehouse Quality Methodology (DWQ)	Jeusfeld et al. 1998 [41]	Data warehousing	16	Correctness, Completeness, Minimality, Traceability, Interpretability, Metadata Evolution, Accessibility (System, Transactional, Security), Usefulness (Interpretability), Timeliness (Currency, Volatility), Responsiveness, Completeness, Credibility, Accuracy, Consistency, Interpretability
3	Total Information Quality Management (TIQM)	English 1999 [42]	Data warehousing	16	Εγγενείς διαστάσεις: Definition conformance (consistency), Completeness, Business rules conformance, Accuracy (to surrogate source), Accuracy (to reality), Precision, Nonduplication, Equivalence of redundant data, Concurrency of redundant data. Πραγματιστικές διαστάσεις: Accessibility, Timeliness, contextual clarity, Derivation integrity, Usability, Rightness (fact completeness), Cost
4	A methodology for information quality assessment (AIMQ)	Lee et al. 2002 [43]	General	14	Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of operation, Freedom from errors, Interpretability, Objectivity, Relevancy, Reputation, Security, Timeliness, Understandability
5	Data Quality Assessment (DQA)	Pipino et al. 2002 [45]	General	16	Accessibility, Appropriate amount of data, Believability, Completeness, Freedom from errors, Consistency, Concise Representation, Relevance, Ease of manipulation, Interpretability, Objectivity, Reputation, Security, Timeliness, Understandability, Value added
6	Information Quality Measurement (IQM)	Eppler and Munzenmaier 2002 [46]	Software / Web Applications	16	Accessibility, Consistency, Timeliness, Conciseness, Maintainability, Currency, Applicability, Convenience, Speed, Comprehensiveness, Clarity, Accuracy, Traceability, Security, Correctness, Interactivity
7	ISTAT methodology (ISTAT)	Falorsi et al. 2003 [48]	General	3	Accuracy, Completeness, Consistency
8	Activity-based Measuring and Evaluating of product information Quality methodology (AMEQ)	Su and Jin 2004 [49]	Construction Industry	29	Consistent representation, Interpretability, Case of understanding, Concise representation, Timeliness, Completeness Value added, Relevance, Appropriateness, Meaningfulness, Lack of confusion, Arrangement, Readable, Reasonability, Precision, Reliability, Freedom from bias, Data Deficiency, Design Deficiency, Operation, Deficiencies, Accuracy, Cost, Objectivity, Believability, Reputation, Accessibility, Correctness, Unambiguity, Consistency
9	Cost-effect Of Low Data Quality (COLDQ)	Loshin 2004 [50]	General	36	**Σχήμα**: Clarity of definition, Comprehensiveness, Flexibility, Robustness, Essentialness, Attribute granularity, Precision of domains, Homogeneity, Identifiability, Obtainability, Relevance, Simplicity/Complexity, Semantic consistency, Syntactic consistency **Δεδομένα**: Accuracy, Null Values, Completeness, Consistency, Currency, Timeliness, Agreement of Usage, Stewardship, Ubiquity **Παρουσίαση**: Appropriateness, Correct Interpretation, Flexibility, Format precision, Portability, Consistency, Use of storage. Πολιτική πληροφοριών: Accessibility, Metadata, Privacy, Security, Redundancy, Cost

10	Data Quality in Cooperative Information Systems (DaQuinCIS)	Scannapieco et al. 2004 [51]	General	5	Accuracy, Completeness, Consistency, Currency, Trustworthiness
11	Methodology for the Quality Assessment of Financial Data (QAFD)	De Amicis and Batini 2004 [52]	Finance	5	Syntactic/Semantic accuracy, Internal/External consistency, Completeness, Currency, Uniqueness
12	Canadian Institute for Health Information methodology (CIHI)	Long and Seko 2005 [53]	Health Management	5	Accuracy, Timeliness, Comparability, Usability, Relevance
13	Comprehensive methodology for Data Quality management (CDQ)	Batini and Scannapieco 2006 [54]	General	17	**Σχήμα**: Correctness with respect to the model, Correctness with respect to Requirements, Completeness, Pertinence, Readability, Normalization **Δεδομένα**: Syntactic/Semantic Accuracy, Semantic Accuracy, Completeness, Consistency, Currency, Timeliness, Volatility, Completability, Reputation, Accessibility, Cost
14	Heterogenous Data Quality Methodology (HDQM)	Batini et al. 2011 [56]	General	2	Accuracy, Currency
15	Capturing Data Quality Requirements for Web Applications by means of DQ_WebRE	César Guerra-García et al. 2012 [98]	Software / Web Applications	15	Accuracy, Completeness, Consistency, Credibility, Currentness, Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability, Understandability, Availability, Portability, Recoverability
16	Information Quality Evaluation Framework: Extending ISO 25012 Data Quality Model	Irfan Rafique et al. 2012 [99]	Software / Web Applications	8	Information accuracy, Information accessibility, Information appropriateness, Efficiency, Confidentiality, Availability, Portability, Recoverability.
17	Big Data quality evaluation across the Big Data value chain	Serhani, M.A., El Kassabi, H.T., Taleb, I. and Nujum, A. 2016 [100]	Big Data	3	Accuracy, Completeness, Consistency
18	Big Data Quality - A Quality Dimensions Evaluation	Taleb, I. et al. 2016 [101]	Big Data	3	Accuracy, Completeness, Consistency
19	Big Data Pre-Processing: Closing the Data Quality Enforcement Loop	Taleb, I. and Serhani, M.A. 2017 [102]	Big Data	3	Accuracy, Completeness, Consistency
20	Context-aware data quality assessment for big data	Ardagna, D., Cappiello, C., Samá, W. and Vitali, M. 2018 [103]	Big Data	4	Accuracy, Completeness, Consistency, Distinctness, Precision, Timeliness, Volume
21	Data Quality Assessment for On-line Monitoring and Measuring System of Power Quality Based on Big Data and Data Provenance Theory	Tian et al. 2018 [104]	Big Data	6	Redundancy, Integrity, Accuracy, Consistency, Timeliness, Intelligence
22	Structured data preparation pipeline for machine learning-applications in production	Frye et al. 2020 [105]	Construction Industry	5	Accuracy, Uniformity, Completeness, Consistency, Currentness
23	An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study	Bekar et al. 2020 [106]	Health Management	6	Accuracy, Completeness, Timeliness, Consistent representation, Accessibility, Relevancy
24	Data-driven methodology for state detection of gearbox in phm context	Chen et al. 2021 [107]	Health Management	4	Integrity, Consistency, Accuracy, Timeliness
25	Data quality certification using ISO/IEC 25012: Industrial experiences	Gualo et al. 2021 [37]	Construction Industry	5	Accuracy, Completeness, Consistency, Credibility, Currentness
26	Method for Data Quality Assessment of Synthetic Industrial Data	Iantovics and Enăchescu 2022 [108]	Construction Industry	3	Sensitivity, Specificity, Accuracy

27	Data quality evaluation for smart multi-sensor process monitoring using data fusion and machine learning algorithms	Segreto and Teti 2022 [109]	Construction Industry	3	Accuracy, Precision, Recall
28	A Data Quality Assessment and Control Method in Multiple Products Manufacturing Process	Xu et al. 2022 [110]	Construction Industry	6	Free-of-error, Appropriate amount of data, Ease of manipulation, Relevancy, Imbalance level, Weighted average
29	An ERP Data Quality Assessment Framework for the Implementation of an APS system using Bayesian Networks	Herrmann et al. 2022 [111]	Software / Web Applications	4	Consistency, Completeness, Appropriate amount of data, Accuracy
30	Data quality assessment and analysis for pest identification in smart agriculture	Yang et al. 2022 [112]	Environment	1	Accuracy
31	Virtual sensors for erroneous data repair in manufacturing a machine learning pipeline	Sen et al. 2023 [113]	Construction Industry	1	Accuracy
32	ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design (DAQUAVORD)	César Guerra-García et al. 2023 [114]	Information Systems	15	Accuracy, Completeness, Consistency, Credibility, Currentness. Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability, Understandability, Availability, Portability, Recoverability
33	An Innovative Deep Architecture for Flight Safety Risk Assessment Based on Time Series Data	Sun et al. 2023 [115]	Aviation	3	Accuracy, Precision, Recall
34	A LOF-IDW based data cleaning method for quality assessment in intelligent compaction of soils	Yao et al. 2023 [116]	Geology	1	Accuracy
35	Machine Learning Model to Enhance the Quality of Software Development Risk Management	Mohamed Ahmed Hamada 2024 [117]	Software / Web Applications	3	Accuracy, Precision, Recall
36	Enhancing SVM Classification of Meningitis with Feature-Adaptive Adagrad in CSF Analysis	Sathiya et al. 2024 [118]	Health Management	3	Accuracy, Sensitivity, Specificity
37	Bayesian Optimization based Support Vector Machine for the Early Warming of Enterprise Financial Risk Analysis	Q. Meng 2024 [119]	Finance	4	Profitability, Asset quality, Debt risk, Operation growth
38	Optimizing Data Quality in Deep Learning through Advanced Analytics	Zhang et al. 2024 [120]	Deep learning	4	Accuracy, Precision, Recall, F1-score or mean squared error
39	Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare	Valevik et al. 2024 [121]	Health Management	5	Similarity, Usability, Privacy, Fairness, Carbon footprint
40	Data quality assessment of interventional trials in public trial databases	R. Iken et al. 2024 [122]	Health Management	4	Consistency, Accuracy, Completeness, Timeliness

41	Data quality in environmental assessment methods - Implications for the operational management in manufacturing	Elsner et al. 2024 [123]	Environment	10	Accuracy, Completeness, Traceability, Attributability, Meta data, Consistency, Time resolution, Compliance, Precision, Error estiamtion
42	Data-oriented QMOOD model for quality assessment of multi-client software applications	Yusuf Özçevik 2024 [124]	Software / Web Applications	6	Reusability, Flexibility, Understandability, Functionality, Extensibility, Effectiveness
43	Assessment of living quality in Guangdong: A hybrid knowledge-based and data-driven approach	Xin-Hui et al. 2024 [125]	Environment	3	Air, Water, Radiation

Πίνακας 4.3 Μέθοδοι και μελέτες ανάλυσης ποιότητας δεδομένων βιβλιογραφικής ανασκόπησης και σχετικά στοιχεία.

Α/Α	Μέθοδος / Μελέτη	Τύπος δεδομένων	Τύπος στρατηγικής (Data-driven ή Process-driven)	Τεχνικές	Μέθοδοι ανάλυσης & Αλγόριθμοι	Εκτίμηση κόστους
1	Total Data Quality Management (TDQM)	S / SM	Process-driven	Process redesign	N/A	NAI
2	The Datawarehouse Quality Methodology (DWQ)	S	Data-driven	Data and schema integration	N/A	NAI
3	Total Information Quality Management (TIQM)	S / SM	Data-driven / Process-driven	Data cleaning, Normalization, Error localization and correction / Process redesign	N/A	NAI
4	A methodology for information quality assessment (AIMQ)	S / SM	N/A	N/A	N/A	OXI
5	Data Quality Assessment (DQA)	S	N/A	N/A	N/A	OXI
6	Information Quality Measurement (IQM)	S / SM	N/A	N/A	N/A	OXI
7	ISTAT methodology (ISTAT)	S / SM	Data-driven / Process-driven	Normalization, Record linkage / Process redesign	N/A	OXI
8	Activity-based Measuring and Evaluating of product information Quality methodology (AMEQ)	S / SM	N/A	N/A	N/A	OXI
9	Cost-effect Of Low Data Quality (COLDQ)	S / SM	Data-driven / Process-driven	Cost optimization / Process control, Process redesign	N/A	NAI
10	Data Quality in Cooperative Information Systems (DaQuinCIS)	S / SM	Data-driven	Source trustworthiness, Record linkage	N/A	OXI
11	Methodology for the Quality Assessment of Financial Data (QAFD)	S	N/A	N/A	N/A	OXI
12	Canadian Institute for Health Information methodology (CIHI)	S / SM	Data-driven	N/A	Statistical algorithms	OXI
13	Comprehensive methodology for Data Quality management (CDQ)	S / SM / US	Data-driven / Process-driven	Normalization, Record linkage, Data and schema integration, Error localization and correction / Process control, Process redesign	N/A	NAI
14	Heterogeneous Data Quality Methodology (HDQM)	S / SM / US	Data-driven / Process-driven	Source improvement, Record linkage / Process control Process redesign	N/A	NAI
15	Capturing Data Quality Requirements for Web Applications by means of DQ_WebRE	S / SM	Data-driven	N/A	Model Driven Architecture - Model Driven Web Engineering	OXI

16	Information Quality Evaluation Framework: Extending ISO 25012 Data Quality Model	S / SM	Data-driven / Process-driven	N/A	ISO 25012 based InQ framework	OXI
17	Big Data quality evaluation across the Big Data value chain	S / SM	Data-driven / Process-driven	Data cleaning, Transformation, Approximation, Filtering, Statistical algorithms / Process control	Paper proposed Algorithms	OXI
18	Big Data Quality - A Quality Dimensions Evaluation	S / SM	Data-driven	N/A	Statistical algorithms	OXI
19	Big Data Pre-Processing: Closing the Data Quality Enforcement Loop	S / SM	Data-driven / Process-driven	Data cleaning, Filtering, Normalization / Process control	Paper proposed Algorithms	OXI
20	Context-aware data quality assessment for big data	S / SM	Data-driven	N/A	SVR method	NAI
21	Data Quality Assessment for On-line Monitoring and Measuring System of Power Quality Based on Big Data and Data Provenance Theory	S / SM	Data-driven	Data cleaning, Filtering	N/A	OXI
22	Structured data preparation pipeline for machine learning-applications in production	S / SM	Data-driven	Data cleaning, Integration & Synchronization, Reduction, Transformation, Augmentation & Balancing	Machine Learning methods	OXI
23	An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study	S / SM	Data-driven	N/A	Machine Learning methods	OXI
24	Data-driven methodology for state detection of gearbox in phm context	S / SM	Data-driven	N/A	Statistical algorithms	OXI
25	Data quality certification using ISO/IEC 25012: Industrial experiences	S / SM	Process-driven	ISO/IEC 25012, 25024, 25040	N/AData Quality Analytic Model based on Data Provenance Relationship	OXI
26	Method for Data Quality Assessment of Synthetic Industrial Data	S / SM	Data-driven	N/A	Statistical algorithms	OXI
27	Data quality evaluation for smart multi-sensor process monitoring using data fusion and machine learning algorithms	S / SM	Data-driven	Sensor monitoring, Innovative signal processing technologies, Data interoperability, Data fusion	Machine Learning methods	OXI
28	A Data Quality Assessment and Control Method in Multiple Products Manufacturing Process	S / SM / US	Data-driven	N/A	Data-centric AI	OXI
29	An ERP Data Quality Assessment Framework for the Implementation of an APS system using Bayesian Networks	S	N/A	N/A	Bayesian Networks	OXI
30	Data quality assessment and analysis for pest identification in smart agriculture	S / SM / US	Data-driven	N/A	Proto-DE, Bound-DE, Multi-Branch	OXI
31	Virtual sensors for erroneous data repair in manufacturing a machine learning pipeline	S / SM / US	Data-driven	Data profiling, Cleaning, Feature engineering, Splitting	Machine Learning methods	OXI

32	ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design (DAQUAVORD)	S / SM / US	Process-driven	ISO/IEC 25012	ISO/IEC 25012	OXI
33	An Innovative Deep Architecture for Flight Safety Risk Assessment Based on Time Series Data	S	Data-driven	N/A	Statistical algorithms	OXI
34	A LOF-IDW based data cleaning method for quality assessment in intelligent compaction of soils	S	Data-driven	Data cleaning (LOF-IDW based method)	lof	OXI
35	Machine Learning Model to Enhance the Quality of Software Development Risk Management	S / SM	N/A	N/A	Machine Learning methods	OXI
36	Enhancing SVM Classification of Meningitis with Feature-Adaptive Adagrad in CSF Analysis	S / SM	Data-driven	N/A	Statistical algorithms	OXI
37	Bayesian Optimization based Support Vector Machine for the Early Warming of Enterprise Financial Risk Analysis	S	Data-driven	N/A	BO-SVM	NAI (Financial risk)
38	Optimizing Data Quality in Deep Learning through Advanced Analytics	S / SM / US	Data-driven	Backtranslation, Word embeddings, Outlier detection, Duplicate removal, Error correction, Synthetic Minority Over-sampling Technique (SMOTE), Data Augmentation	Statistical algorithms	OXI
39	Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare	S / SM	Data-driven	N/A	Statistical algorithms	OXI
40	Data quality assessment of interventional trials in public trial databases	SM	Data-driven	N/A	IBM SPSS Statistics	OXI
41	Data quality in environmental assessment methods - Implications for the operational management in manufacturing	S / SM / US	Process-driven	Process control	N/A	OXI
42	Data-oriented QMOOD model for quality assessment of multi-client software applications	S / SM	Data-driven	Data-oriented QMOOD	N/A	OXI
43	Assessment of living quality in Guangdong: A hybrid knowledge-based and data-driven approach	S	Data-driven	N/A	Clustering (HCA, k-means, FCM, DBSCAN)	OXI

Κεφάλαιο 5ο – Λογισμικά ανάλυσης και ποιότητας δεδομένων

Ενότητα 5.1 – Εισαγωγή

Πλήθος λογισμικών ανάλυσης και ποιότητας δεδομένων έχουν αναπτυχθεί και προσφερθεί στην αγορά για να βοηθήσουν τους οργανισμούς και τις εταιρείες, ώστε να βελτιώσουν τις δομές δεδομένων τους. Παρακάτω παρουσιάζονται 10 από τα πιο γνωστά εμπορικά προϊόντα και καταγράφονται ορισμένα χαρακτηριστικά και λειτουργίες αυτών (και με τη βοήθεια των βιβλιογραφικών αναφορών 126 και 127):

1. IBM InfoSphere Information Server for Data Quality
2. SAP Master Data Governance
3. Talend Data Quality
4. Ataccama Data Quality & Governance
5. Informatica Cloud Data Quality
6. Oracle Enterprise Data Quality
7. Melissa Unison
8. Precisely Data Integrity Suite
9. SAS Data Quality ή Viya
10. Collibra Data Quality & Observability
11. Experian Aperture Data Studio

Ενότητα 5.2 – Σύγχρονα λογισμικά ποιότητας δεδομένων

1. IBM InfoSphere Information Server for Data Quality

Η πλατφόρμα IBM InfoSphere Information Server [128] βοηθάει στη βελτιστοποίηση της διαχείρισης και ποιότητας των δεδομένων, μετατρέποντας ανεπεξέργαστα δεδομένα σε αξιόπιστες πληροφορίες. Αναλύει και παρακολουθεί συνεχώς (monitoring) την ποιότητα των δεδομένων, καθαρίζει, τυποποιεί δεδομένα και ταιριάζει τις εγγραφές για την εξάλειψη των διπλοτύπων.

Η πλατφόρμα IBM InfoSphere Information Server παρέχει λειτουργίες καθαρισμού δεδομένων που αυτοματοποιούν τη διερεύνηση των δεδομένων προέλευσης, επιτρέποντας την τυποποίηση πληροφοριών και την αντιστοίχιση αρχείων με βάση καθορισμένους επιχειρηματικούς κανόνες. Υποστηρίζει, επίσης, τη συνεχή παρακολούθηση της ποιότητας για να μειώσει τη διάδοση εσφαλμένων ή ασυνεπών δεδομένων. Η πλατφόρμα επιτρέπει τη λειτουργία σε οποιαδήποτε θέση εγκατάστασης, είτε πρόκειται για εσωτερική εγκατάσταση, είτε στο cloud ή συνδυασμό και των δύο. Μία από τις βασικές λειτουργίες της πλατφόρμας είναι η διαχείριση ζητημάτων ποιότητας δεδομένων, όπου πραγματοποιείται η δημιουργία ενός σχεδίου διόρθωσης μέσω μετρήσεων, σύμφωνα με τους καθορισμένους επιχειρηματικούς στόχους. Βοηθά στη διαχείριση ενός προγράμματος διακυβέρνησης δεδομένων και επιτρέπει την προσαρμογή των διαδικασιών τυποποίησης δεδομένων, σύμφωνα με τις επιχειρηματικές απαιτήσεις, όπως ο εμπλουτισμός ή ο καθαρισμός δεδομένων. Επίσης, περιλαμβάνει μια λειτουργία "ταξινόμησης" που προσδιορίζει τη θέση των Προσωπικών Πληροφοριών Αναγνώρισης (Personally Identifiable Information / PII).

Η πλατφόρμα IBM InfoSphere Information Server παρέχει ισχυρές δυνατότητες διαχείρισης ποιότητας δεδομένων, με ενσωματωμένα εργαλεία που βοηθούν στη διατήρηση του απορρήτου σε ένα σύνολο δεδομένων.

2. SAP Master Data Governance

Το SAP Master Data Governance [129] είναι ένας κεντρικός κόμβος/σύστημα για τη διαχείριση και βελτίωση της ποιότητας των κρίσιμων για την επιχείρηση δεδομένων, επιτρέποντας πιο αποτελεσματικές πρακτικές εργασίας, ώστε να οδηγούν σε βελτιωμένες διαδικασίες λήψης αποφάσεων. Το SAP Master Data Governance προσφέρει στο χρήστη, τη δυνατότητα συγκέντρωσης των κύριων δεδομένων και την κεντρική διαχείρισής τους, χρησιμοποιώντας κατάλληλες στρώσεις δεδομένων (data layers).

Το SAP Master Data Governance προσφέρει διαχείριση δεδομένων για συγκεκριμένους τομείς (π.χ. πρώτες ύλες, προϊόντα, πελάτες). Με αυτόν τον τρόπο, οι επιχειρήσεις μπορούν να ελέγχουν και να ενοποιούν ή να δημιουργούν, να αλλάζουν και να διανέμουν κύρια δεδομένα στα εταιρικά τους συστήματα. Η αυστηρή ενοποίηση

με άλλες λύσεις SAP υποστηρίζει την επαναχρησιμοποίηση μοντέλων δεδομένων, επιχειρηματικής λογικής και πλαισίων επικύρωσης. Η εφαρμογή υποστηρίζει επίσης ανοιχτή ενοποίηση με προϊόντα και υπηρεσίες τρίτων. Το SAP Data Governance δίνει τη δυνατότητα στις ομάδες να κατέχουν μοναδικά κύρια χαρακτηριστικά δεδομένων και διατηρεί επικυρωμένες τιμές για συγκεκριμένα σημεία δεδομένων, μέσω δρομολόγησης και ειδοποίησης της ροής εργασίας. Για την ποιότητα των δεδομένων και την ανάλυση διεργασιών, ορίζει, επικυρώνει και παρακολουθεί τους επιχειρηματικούς κανόνες, επιβεβαιώνοντας την αναγνωσιμότητα στα κύρια δεδομένα και αναλύοντας τις επιδόσεις διαχείρισης.

Το SAP Master Data Governance μπορεί να λειτουργεί είτε εντός των εγκαταστάσεων μιας εταιρείας ή οργανισμού είτε μέσω cloud για υβριδικά και cloud συστήματα. Υποστηρίζει όλους τους κύριους τομείς δεδομένων και έχει προκατασκευασμένα μοντέλα δεδομένων, επιχειρηματικούς κανόνες, ροές εργασίας και διεπαφές χρήστη.

3. Talend Data Fabric

Το Talend's Data Fabric [130] είναι μια ολοκληρωμένη πλατφόρμα που ενσωματώνει την ποιότητα, την ακεραιότητα και τη διακυβέρνηση δεδομένων σε ένα σύστημα. Η ενότητα ενσωμάτωσης δεδομένων διευκολύνει τη συλλογή, τον μετασχηματισμό και τη χαρτογράφηση δεδομένων. Διασφαλίζει την εμπιστοσύνη τους καθ' όλη τη διάρκεια του κύκλου ζωής των δεδομένων, με την ακεραιότητα των δεδομένων και τα χαρακτηριστικά διακυβέρνησής τους. Επιπρόσθετα, η μονάδα ποιότητας δεδομένων καθαρίζει και αποκρύπτει δεδομένα σε πραγματικό χρόνο. Το Talend Data Fabric είναι επίσης εξοπλισμένο για εφαρμογή και ολοκλήρωση API, παρέχοντας στους χρήστες τη δυνατότητα να μοιράζονται και να προσδίδουν αξία από αξιόπιστα δεδομένα στο εσωτερικό και εξωτερικό περιβάλλον ενός οργανισμού.

Κύριο συστατικό του Talend Data Fabric, είναι η ενότητα Data Quality, η οποία χρησιμοποιεί μεθόδους μηχανικής μάθησης, ώστε να προτείνει λύσεις για ζητήματα ποιότητας δεδομένων κατά τη διάρκεια ροής δεδομένων σε πραγματικό χρόνο. Εντός της ενότητας Data Quality, η λειτουργία προφίλ δεδομένων του Talend επιτρέπει τον γρήγορο εντοπισμό ζητημάτων ποιότητας δεδομένων και την ανακάλυψη κρυφών μοτίβων και ανωμαλιών. Αυτό γίνεται δυνατό μέσω συνοπτικών στατιστικών και γραφικών αναπαραστάσεων. Το ενσωματωμένο Talend Trust Score προσφέρει μια

άμεση, κατανοητή και εφαρμόσιμη αξιολόγηση της εμπιστοσύνης των δεδομένων. Αυτή η δυνατότητα διασφαλίζει την ασφαλή κοινή χρήση συνόλων δεδομένων και υποδεικνύει ποια σύνολα δεδομένων πρέπει να υποβληθούν σε πρόσθετη επεξεργασία.

Το Talend επεξεργάζεται αυτόματα τα εισερχόμενα δεδομένα και τα εμπλουτίζει με λεπτομέρειες από εξωτερικές πηγές. Η πλατφόρμα παρέχει δυνατότητα απόκρυψης ευαίσθητων δεδομένων, διασφαλίζοντας την ευθυγράμμιση με εσωτερικούς και εξωτερικούς κανονισμούς απορρήτου και προστασίας δεδομένων.

4. Ataccama Data Quality & Governance

Η πλατφόρμα Ποιότητας και Διακυβέρνησης της Ataccama [131] δεδομένων βοηθά στην εξάλειψη των ασυνεπειών των δεδομένων, στην αύξηση της ακρίβειας και στην αποκατάσταση της εμπιστοσύνης στους πόρους δεδομένων ενός οργανισμού. Στόχος είναι η παροχή ασφαλών, ποιοτικών δεδομένων για αξιόπιστες αναλύσεις και αναφορές και η μείωση των κινδύνων που σχετίζονται με τη διαχείριση δεδομένων, με ενσωματωμένη προστασία των ευαίσθητων δεδομένων.

Δύο καίρια χαρακτηριστικά της πλατφόρμας Ataccama Data Quality & Governance είναι η προετοιμασία και επικύρωση δεδομένων με τη βοήθεια της τεχνητής νοημοσύνης (AI), καθώς και η προληπτική διασφάλιση ποιότητας δεδομένων. Η τεχνολογία τεχνητής νοημοσύνης χρησιμοποιείται για την βελτίωση της αναγνωσιμότητας των δεδομένων, την αυτοματοποίηση των εντατικών διαδικασιών προετοιμασίας και επικύρωσης δεδομένων, την επιτάχυνση των λειτουργιών και την παροχή έγκαιρων, αξιόπιστων πληροφοριών στους διαχειριστές. Η προληπτική λειτουργία ποιότητας δεδομένων περιλαμβάνει παρακολούθηση, δημιουργία προφίλ και ανίχνευση ανωμαλιών. Αυτές οι δυνατότητες λειτουργούν συνεχώς για τον εντοπισμό και τη διόρθωση προβλημάτων σε πραγματικό χρόνο. Χρησιμοποιούν αυτοματοποιημένες ειδοποιήσεις, ώστε οι ομάδες αναλυτών να ενημερώνονται γρήγορα για τυχόν ζητήματα που δεν μπορούν να αντιμετωπιστούν αυτόματα. Η πλατφόρμα της Ataccama ενσωματώνει επίσης τη διακυβέρνηση δεδομένων, συμπεριλαμβανομένης της διαχείρισης μεταδεδομένων, της γενεαλογίας και της διαχείρισης. Χρησιμοποιεί ελέγχους πρόσβασης καθώς και μέτρα ασφαλείας που

αποτρέπουν μη εξουσιοδοτημένες τροποποιήσεις και χειρισμό δεδομένων, παρέχοντας ένα ασφαλές και ελεγχόμενο περιβάλλον για τα δεδομένα.

Συνολικά, η Ataccama είναι μια αξιόπιστη επιλογή για οργανισμούς που προσπαθούν να εξισορροπήσουν την ασφάλεια και τον έλεγχο με την ανάγκη για προσβασιμότητα δεδομένων σε ολόκληρο τον οργανισμό. Ενσωματώνοντας στενά τη διακυβέρνηση και την ποιότητα των δεδομένων, η πλατφόρμα της Ataccama μπορεί να συμβάλει στην ενίσχυση της επιχειρηματικής και λειτουργικής αποτελεσματικότητας.

5. Informatica Cloud Data Quality

Το Informatica Cloud Data Quality [132] αποτελεί μια ολοκληρωμένη λύση που βοηθά τις επιχειρήσεις να εντοπίζουν, να διορθώνουν και να παρακολουθούν ζητήματα ποιότητας δεδομένων στις εφαρμογές τους. Υποστηρίζει μια συνεργατική προσέγγιση, που συνδυάζει τις προσπάθειες των επιχειρηματικών χρηστών και του προσωπικού πληροφορικής για την ανάπτυξη ενός περιβάλλοντος που βασίζεται σε δεδομένα. Αυτή η συνεργασία προωθεί την ταχύτερη υλοποίηση σε τεχνολογία σύννεφου (cloud), μέσω ταχείας μετεγκατάστασης και πληροφοριών υψηλής εμπιστοσύνης από πηγές δεδομένων, όπως αποθήκες δεδομένων cloud, “λίμνες” δεδομένων και εφαρμογές τύπου Λογισμικού ως Υπηρεσία (System as a Service / SaaS).

Βασικό πλεονέκτημα του Informatica Cloud Data Quality αποτελεί το γεγονός ότι επιτρέπει τον εντοπισμό και την επίλυση προβλημάτων ποιότητας δεδομένων, χωρίς την ανάγκη συμπληρωματικού κώδικα Τεχνολογίας Πληροφοριών (Information Technology / IT) ή ανάπτυξής του. Αυτό έχει ως αποτέλεσμα αυξημένη ασφάλεια, αξιοπιστία και εστίαση στη λειτουργική αριστεία, χωρίς πρόσθετες επενδύσεις σε τεχνολογικές υποδομές. Το Informatica Cloud Data Quality περιλαμβάνει επίσης ένα πλούσιο σύνολο μετασχηματισμών ποιότητας δεδομένων και καθολικών συνδέσεων, παρέχοντας ολοκληρωμένη υποστήριξη για όλους τους τύπους δεδομένων και περιπτώσεων χρήσης.

Ένα άλλο σημαντικό χαρακτηριστικό εργαλείο είναι το CLAIRE, καθώς παρέχει τεχνητή νοημοσύνη που βασίζεται σε μετα-δεδομένα για να επιτρέψει έξυπνες συστάσεις κανόνων ποιότητας δεδομένων που προέρχονται από παρόμοια μοτίβα διαχείρισης δεδομένων. Κατά συνέπεια, ενισχύει την αυτόματη ανίχνευση ομοιότητας

δεδομένων, για τον εντοπισμό και την εξάλειψη των διπλών καταχωρήσεων δεδομένων.

Συνολικά, το Informatica Cloud Data Quality απλοποιεί τις διοικητικές διαδικασίες και μειώνει το γενικό κόστος, παρέχοντας ένα ενοποιημένο εργαλείο ποιότητας δεδομένων που μπορεί να χρησιμοποιηθεί σε τμήματα, εφαρμογές, ακόμη και μοντέλα ανάπτυξης, όλα βασισμένα σε τεχνολογία σύννεφου.

6. Oracle Enterprise Data Quality (EDQ)

Η οικογένεια προϊόντων Oracle Enterprise Data Quality [133] βοηθά τους οργανισμούς να επιτύχουν τη μέγιστη αξία στις κρίσιμες εφαρμογές τους, παρέχοντας δεδομένα κατάλληλα για τον εκάστοτε σκοπό. Αυτά τα προϊόντα επιτρέπουν επίσης σε άτομα και συνεργατικές ομάδες να εντοπίζουν γρήγορα και εύκολα και να επιλύουν τυχόν προβλήματα στα υποκείμενα δεδομένα.

Το EDQ είναι η συνολική λύση της Oracle για διακυβέρνηση δεδομένων και διαχείριση ποιότητας δεδομένων. Ως μέρος της ομάδας προϊόντων Oracle Fusion Middleware, η πλατφόρμα διαθέτει αρκετούς ενσωματωμένους μετασχηματισμούς για τη γρήγορη διαμόρφωση προφίλ δεδομένων και για την ανακάλυψη μοτίβων και προβλημάτων ποιότητας. Το EDQ είναι επίσης πλήρως επεκτάσιμο, προσφέροντας επεκτάσεις για επαλήθευση διεύθυνσης και μπορεί να χρησιμοποιηθεί για επεξεργασία πακέτων δεδομένων σε πραγματικό χρόνο με άλλα εργαλεία ολοκλήρωσης.

Η πλατφόρμα παρέχει ένα ολοκληρωμένο περιβάλλον για την κατανόηση, τη βελτίωση και τη διαχείριση της ποιότητας των δεδομένων σε διάφορες επιχειρηματικές διαδικασίες. Υπάρχει επίσης μια εικονική μηχανή για τη δοκιμή του EDQ με ένα δείγμα δεδομένων, το οποίο μπορεί να χρησιμοποιηθεί για το προφίλ, τον έλεγχο, την τυποποίηση, την κανονικοποίηση και την κατάργηση των διπλοεγγραφών.

Τα προϊόντα Oracle Enterprise Data Quality προσφέρουν τη δυνατότητα βελτίωσης της ποιότητας των δεδομένων σε όλους τους ενδιαφερόμενους και για οποιαδήποτε πρωτοβουλία διαχείρισης δεδομένων. Τα προϊόντα Oracle Enterprise Data Quality προσφέρουν:

- Προφίλ, έλεγχος και πίνακες ελέγχου δεδομένων

- Ανάλυση και τυποποίηση δεδομένων
- Ταίριασμα και συγχώνευση δεδομένων
- Διαχείριση υποθέσεων δεδομένων
- Επαλήθευση διεύθυνσης δεδομένων

7. Melissa Unison

Η πλατφόρμα Unison της Melissa [134] συνδράμει τους επιχειρησιακούς χρήστες με ένα πλούσιο σύνολο εργαλείων μετασχηματισμού ποιότητας δεδομένων, συμπεριλαμβανομένης της τυποποίησης, της επικύρωσης και του εμπλουτισμού δεδομένων για την παροχή πληροφοριών υψηλής ποιότητας σε όλη την επιχείρηση. Η ανάλυση, η αντιστοίχιση, η δημιουργία προφίλ και ο καθαρισμός δεδομένων ενσωματώνονται σε μια τεχνολογία που μπορεί να κλιμακωθεί σε πολλούς διακομιστές και να λειτουργεί εκτός σύνδεσης. Αυτό επιτρέπει την ατομική διαχείριση των δεδομένων για την κάλυψη των απαιτήσεων συμμόρφωσης και ασφάλειας. Πιο συγκεκριμένα, η πλατφόρμα προσφέρει τις εξής δυνατότητες:

- Τη δημιουργία προφίλ δεδομένων, μέσω της χρήσης πρότυπων και ανεπτυγμένων μετα-δεδομένων.
- Την επαλήθευση, εμπλουτισμό και την ενοποίηση των δεδομένων.
- Τον καθαρισμό και την τυποποίηση των δεδομένων, μέσω μεθόδων μηχανικής μάθησης.
- Τον καθαρισμό διπλότυπων εγγραφών.
- Την αναζήτηση και εύρεση τάσεων των δεδομένων με τη βοήθεια γραφημάτων και αναφορών.
- Παροχή περιορισμών πρόσβασης σε επίπεδο χρήστη, ασφάλειας διαχείρισης δεδομένων τοπικής εγκατάστασης και λεπτομερή καταγραφή για καλύτερο έλεγχο, διασφαλίζοντας τη συμμόρφωση με τη σχετική νομοθεσία.

8. Precisely Data Integrity Suite

Το Precisely Data Integrity Suite [135] επιτρέπει ευρεία διαχείριση των δεδομένων, προσφέροντας μια σουίτα αρκετών διαλειτουργικών υπηρεσιών. Προσφέρει εργαλεία για τη δημιουργία προφίλ δεδομένων, τον καθαρισμό, την τυποποίηση και την αντιστοίχιση μεταξύ διαφόρων πηγών δεδομένων, αντιμετωπίζοντας το ζήτημα της ποιότητας δεδομένων ακόμα και σε πραγματικό χρόνο. Προσφέρεται η δυνατότητα οπτικοποίησης των αλλαγών δεδομένων.

Το Precisely Data Integrity Suite συνδυάζοντας λειτουργίες για την ποιότητα, την επικύρωση, την ανακάλυψη «ανωμαλιών» στα δεδομένα, τη διακυβέρνηση, τη γεωκωδικοποίηση, τον εμπλουτισμό και την εύρεση προτύπων και τάσεων στα δεδομένα, αποτελεί μια ενδιαφέρουσα επιλογή για τη διαχείριση των δεδομένων.

Μερικά από τα οφέλη της χρήσης του είναι:

- Βελτιωμένη απόδοση με τυποποιημένα και επαναλαμβανόμενα μέτρα ποιότητας δεδομένων.
- Αυξημένη εμπιστοσύνη στις αποφάσεις με ακριβή και αξιόπιστα δεδομένα.
- Μειωμένο κόστος με τη μη αποθήκευση και διαχείριση πολλαπλών αντιγράφων δεδομένων.
- Επικαιροποίηση των αποφάσεων μέσω του εμπλουτισμού των δεδομένων.
- Μείωση χρόνου εργασίας, ώστε τα δεδομένα να είναι έτοιμα για λειτουργίες και αναλύσεις.

Σημειώνεται ότι η πλατφόρμα μπορεί να χρησιμοποιηθεί στο cloud, σε hybrid-cloud ή σε εσωτερικές εγκαταστάσεις του χρήστη.

9. SAS Data Quality ή SAS Viya

Το SAS Data Quality [136] είναι μια λύση λογισμικού σχεδιασμένη για να βελτιώνει την ακρίβεια, τη συνέπεια και την πληρότητα των δεδομένων. Η πλατφόρμα παρέχει ολοκληρωμένα εργαλεία για την ανάλυση της δομής των δεδομένων, τον εντοπισμό ασυνεπειών και την εφαρμογή μετασχηματισμών για τη διόρθωση σφαλμάτων μορφοποίησης ή ελλιπών και ακραίων τιμών. Βοηθάει στον εντοπισμό και τη συγχώνευση διπλότυπων εγγραφών, ενώ επίσης τυποποιεί μορφές δεδομένων (π.χ.

ημερομηνίες, διευθύνσεις) για να διασφαλίσει τη συνέπεια μεταξύ διαφορετικών πηγών. Παρέχει μια έτοιμη βιβλιοθήκη δεδομένων αναφοράς που ονομάζεται SAS Quality Knowledge Base (QKB), η οποία βοηθά στην αυτοματοποίηση κοινών εργασιών ποιότητας δεδομένων για εξοικονόμηση χρόνου. Το SAS Data Quality χρησιμοποιεί τεχνικές της τεχνητής νοημοσύνης στη βελτίωση της ποιότητας των δεδομένων, ενώ προσφέρει διάφορα πακέτα (πχ προβλέψων, τεχνικών μηχανικής μάθησης κ.α.), ανάλογα με τις ανάγκες και τις οικονομικές δυνατότητες των πελατών.

Σημειώνεται ότι η πλατφόρμα μπορεί να χρησιμοποιηθεί στο cloud ή σε εσωτερικές εγκαταστάσεις του χρήστη.

10. Collibra Data Quality & Observability

Η εταιρεία Collibra διαθέτει στο εμπόριο μια πλατφόρμα (Data Intelligence Platform) πολλών διαφορετικών λειτουργιών, όσον αφορά τη διαχείριση των δεδομένων. Αυτά είναι η διακυβέρνηση δεδομένων με τη βοήθεια τεχνητής νοημοσύνης, η ιδιωτικότητα δεδομένων και η ασφάλεια, ο εμπλουτισμός των δεδομένων και η βελτίωση της ποιότητας των δεδομένων μέσω της πλατφόρμας Data Quality & Observability [137].

Η πλατφόρμα λειτουργεί και μέσω cloud και δίνει τη δυνατότητα για διασύνδεση με πάνω από 40 βάσεις δεδομένων (π.χ. Oracle, MySQL). Η πλατφόρμα χρησιμοποιεί ελέγχους παρακολούθησης των δεδομένων και των τάσεών τους, μέσω προσαρμοσμένων κανόνων με τη βοήθεια τεχνητής νοημοσύνης. Για τη διόρθωση των λανθασμένων εγγραφών, εφαρμόζεται ταξινόμησης βάσει εξελιγμένων κανόνων ποιότητας. Οι κανόνες ποιότητας μπορούν να αναπτυχθούν μέσω τεχνητής νοημοσύνης σε διατιθέμενο προγραμματιστικό περιβάλλον. Ο έλεγχος και η παρακολούθηση της ποιότητας των δεδομένων δίνουν τη δυνατότητα για την έκδοση ορθών πινάκων δεδομένων και αναλύσεων και αναφορών.

Ενότητα 5.3 – Σύγκριση σύγχρονων λογισμικών ποιότητας δεδομένων

Σε συνέχεια της παράθεσης των παραπάνω λογισμικών ποιότητας δεδομένων, παρατίθενται μια συνοπτική σύγκριση μεταξύ τους, με τη βοήθεια του παρακάτω πίνακα (Πίνακας 5.1):

Πίνακας 5.1 Συγκριτική ανάλυση λογισμικών ποιότητας δεδομένων.

Α/Α	Λογισμικό	Εταιρεία	Τύπος πλατφόρμας	Μηχανική μάθηση	Τεχνητή νοημοσύνη (AI)	Παρακολούθηση (Monitoring)	Έλεγχος & Βελτιστοποίηση διαδικασιών	Έλεγχος κόστους
1	InfoSphere Information Server for Data Quality	IBM	+Cloud	N/A	OXI	NAI	NAI	OXI
2	Master Data Governance	SAP	+Cloud	N/A	OXI	NAI	NAI	OXI
3	Data Fabric	Talend	+Cloud	NAI	OXI	NAI	OXI	OXI
4	Data Quality & Governance	Ataccama	+Cloud	N/A	NAI	NAI	NAI	OXI
5	Cloud Data Quality	Informatica	+Cloud	N/A	NAI	NAI	OXI	OXI
6	Enterprise Data Quality	Oracle	+Cloud	N/A	OXI	NAI	OXI	OXI
7	Unison	Melissa	+Cloud	NAI	OXI	NAI	OXI	OXI
8	Data Integrity Suite	Precisely	+Cloud	N/A	OXI	NAI	OXI	OXI
9	Data Quality ή Viya	SAS	+Cloud	NAI	NAI	NAI	OXI	OXI
10	Data Quality & Observability	Collibra	+Cloud	N/A	NAI	NAI	OXI	OXI

Βάσει του παραπάνω Πίνακα 5.1, παρουσιάζονται στοιχεία των λογισμικών, ώστε να εντοπιστούν διαφοροποιήσεις. Είναι εμφανές, ότι πλέον όλα τα λογισμικά λειτουργούν ως πλατφόρμες στο cloud, ώστε να διατίθενται δυνατότητες ταχύτατης, αξιόπιστης και ταυτόχρονης μεταφοράς δεδομένων και αναλύσεων στις διάφορες δομές ενός οργανισμού ή επιχείρησης. Η τεχνολογία σύννεφου είναι διαδραστική και παρέχει δυνατότητες για αξιοποίηση μεγάλων βάσεων δεδομένων (πχ. Oracle) οι οποίες είναι συμβατές με τα παραπάνω λογισμικά.

Οι εν λόγω πλατφόρμες πέρα από την ποιότητα δεδομένων, συνήθως εμπορεύονται και συμπληρωματικά πακέτα υψηλών δυνατοτήτων και λειτουργιών, όπως η διακυβέρνηση, η γεωκωδικοποίηση και ο εμπλουτισμός δεδομένων.

Ορισμένα από τα λογισμικά στηρίζουν τις αναλύσεις τους σε τεχνικές και αλγορίθμους μηχανικής μάθησης, ενώ άλλα λογισμικά έχουν πλέον εστιάσει στις δυνατότητες της τεχνητής νοημοσύνης. Στον Πίνακα 5.1, παρουσιάζονται τα λογισμικά που στηρίζονται σε κάθε μέθοδο, ενώ σε άλλα δεν είναι δημοσιευμένες οι μέθοδοι που εφαρμόζονται.

Σε όλα τα λογισμικά υπάρχει η δυνατότητα για παρακολούθηση της ποιότητας και της τάσης των δεδομένων με τρέχουσες αναφορές και γραφήματα τα οποία παρουσιάζουν

τα στοιχεία που επιθυμεί ο αναλυτής, ώστε να μπορεί άμεσα να προχωρήσει σε τροποποιήσεις και αλλαγές των κανόνων ποιότητας.

Στα υπόψη λογισμικά, ζήτημα παρουσιάζεται στην απουσία ύπαρξης ελέγχου του κόστους σε ένα εφαρμοσμένο σχέδιο ποιότητας δεδομένων. Όπως προέκυψε και στο Κεφάλαιο 4, αν και το κόστος των επιχειρήσεων ή οργανισμών, λόγω λανθασμένων δεδομένων αναφέρονται όλο και πιο σύχνα ως μείζον ζήτημα, λίγες μεθοδολογίες πραγματεύονται τη συγκεκριμένη έννοια. Αυτό έχει ως αποτέλεσμα, τα λογισμικά να δίνουν ελάχιστη σημασία στην εφαρμογή αλγόριθμων και τεχνικών, ώστε να εξαχθεί ένα συμπέρασμα για τα κόστη, βάσει του οποίου θα παρθούν σημαντικές αποφάσεις.

Τέλος, λίγα λογισμικά πραγματοποιούν έλεγχο των διαδικασιών, κάτι το οποίο είναι σημαντικό για την προκύπτουσα ποιότητα δεδομένων. Έχει σημειωθεί και σε προηγούμενα κεφάλαια, ότι ο έλεγχος και η αλλαγή των διαδικασιών, αν και επιφέρει σημαντικά οφέλη είναι πιο δύσκολο να εφαρμοστεί σε μια δομή, καθώς απαιτεί τροποποίηση δομών και λειτουργιών, κάτι το οποίο κοστίζει σε χρόνο και χρήμα.

Κεφάλαιο 6ο – Συμπεράσματα εργασίας & προτάσεις για περαιτέρω διερεύνηση

Ενότητα 6.1 – Συμπεράσματα εργασίας

Ανακεφαλαιώνοντας, στην παρούσα εργασία παρουσιάστηκε αρχικώς η έννοια των δεδομένων και της ποιότητας δεδομένων με τις αντίστοιχες διαστάσεις. Στη συνέχεια, παρουσιάστηκαν οι Ευρωπαϊκές Οδηγίες 2021 για το ζήτημα της ποιότητας δεδομένων, αναλύοντας πλήρως τις διαστάσεις που περιλαμβάνονται. Στο τρίτο κεφάλαιο παρουσιάστηκαν πτυχές το προτύπου Ποιότητας Δεδομένων ISO 25000. Στο τέταρτο κεφάλαιο, παρουσιάστηκαν σύγχρονες μεθόδοι ανάλυσης δεδομένων από τη διεθνή βιβλιογραφία, εστιάζοντας τόσο στο ερευνητικό όσο και στο κανονιστικό πεδίο, καθώς και συνοπτική σύγκριση αυτών. Στο πέμπτο κεφάλαιο, επιχειρήθηκε προσπάθεια παρουσίασης των σύγχρονων λογισμικών που εστιάζουν στην ποιότητα των δεδομένων και μια συνοπτική σύγκριση αυτών.

Από τα παραπάνω προκύπτουν βασικά συμπεράσματα, όπως:

- Η αντιμετώπιση του ζητήματος της ποιότητας των δεδομένων είναι πολυσύνθετη, καθώς σημαντικό ρόλο διαδραματίζει ο τύπος των δεδομένων που χρησιμοποιούνται.
- Οι διαστάσεις ποιότητας επιλέγονται ανάλογα με το ζήτημα, καθώς και την δυνατότητα (σε χρόνο και κόστος) ανάλυσης που υπάρχει από τον εκάστοτε ερευνητή, επιχείρηση ή οργανισμό.
- Οι Ευρωπαϊκές Οδηγίες 2021 προσφέρουν το σύστημα διαστάσεων FAIR, το οποίο προσφέρει ένα πλήρες πρότυπο διαστάσεων ποιότητας δεδομένων.
- Το Πρότυπο ISO 25000 προσφέρει ένα ακόμα πρότυπο/πλαίσιο για τον έλεγχο, και την αξιολόγηση της ποιότητας δεδομένων, αλλά και των αντίστοιχων διαδικασιών.
- Η διεθνής βιβλιογραφία προσφέρει πληθώρα μεθοδολογιών και μεθόδων ανά τομέα έρευνας για την ανάλυση, αξιολόγηση και βελτίωση της ποιότητας των δεδομένων. Λιγότερες, ωστόσο, είναι οι μελέτες για τη βελτίωση των διαδικασιών αλλά και του σχετικού κόστους.
- Πλήθος λογισμικών συνδράμουν στην ανάλυση, αξιολόγηση και βελτίωση της ποιότητας των δεδομένων, με χρήση μεθόδων μηχανικής μάθησης και τεχνητής

νοημοσύνης. Τα πακέτα λογισμικού προσφέρουν πλήθος διαλειτουργικών δυνατοτήτων, όπως η κυβερνησιμότητα και ο εμπλουτισμός δεδομένων και μέσω της τεχνολογίας σύννεφου.

Ενότητα 6.2 – Προτάσεις για περαιτέρω διερεύνηση

Καθώς το ζήτημα της ποιότητας δεδομένων είναι πολύπλευρο και η έρευνά του εξελίσσεται ραγδαία τις τελευταίες δεκαετίες, υπάρχουν μεγάλα περιθώρια για περαιτέρω αναζήτηση σε αρκετούς τομείς.

Σημαντική θα ήταν η περαιτέρω έρευνα και εστίαση σε μελέτες που αφορούν το Πρότυπο ISO 25000, ώστε να αναδειχθούν καλύτερα τα πλεονεκτήματα ή/και οι ατέλειες του εν λόγω προτύπου στην πράξη.

Επιπρόσθετα, σημαντική θα ήταν η εστίαση σε μεθοδολογίες που αφορούν τον έλεγχο, την αξιολόγηση και την βελτιστοποίηση των διαδικασιών ποιότητας δεδομένων, καθώς και πιθανά λογισμικά που να εξετάζουν σχετικά θέματα, ώστε να αναδειχθεί περαιτέρω η σημασία των διαδικασιών ποιότητας δεδομένων στην επίλυση καίριων ζητημάτων.

Τελός, η εστίαση σε μελέτες που ασχολούνται με το κόστος της κακής ποιότητας δεδομένων, καθώς και τη σχέση μεταξύ κόστους και μεταβολής διαδικασιών (έστω και σε συγκεκριμένο τομέα ζητήματος), θα μπορούσε να προσφέρει σημαντικα συμπεράσματα.

Βιβλιογραφικές Αναφορές

1. D. Reinsel, J. Gantz, J. Rydning, Data Age 2025: The Evolution of Data to Life Critical. MA, USA: An IDC White Paper. International Data Corporation (IDC), 2017.
2. International Organization for Standardization, Geneva, Switzerland: ISO 9000:2015, Quality Management Systems, 2015.
3. T. Redman, Data Quality: The Field Guide. MA, USA: Digital Press: Boston, 2001.
4. R. Wang, D. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manag. Inf. Syst., 1996.
5. B. Kahn, D. Strong, R. Wang, Information quality benchmarks: Product and service performance. Commun. ACM, 2002.
6. C. Fürber, Data Quality Management with Semantic Technologies, 1st ed. Wiesbaden, Germany: Springer Gabler, 2015.
7. L. Piwek, D. Ellis, S. Andrews, A. Joinson, The Rise of Consumer Health Wearables: Promises and Barriers. PLoS Med, 2016.
8. S. Jones, Health & Fitness Wearables: Market Size, Trends & Vendor Strategies. In S. Jones. Hampshire, UK: Juniper Research Ltd.: Basingstoke, 2020.
9. K. Rothman, Epidemiology: An Introduction, 2nd ed.. NY, USA: Oxford University Press: New York, 2012.
10. W.-Y. Loh, Q. Zhang, W. Zhang, P. Zhou, Missing data, imputation and regression trees. Sin, 2020.
11. T. McCausland, The Bad Data Problem. In T. McCausland. Res.-Technol. Manag.. 2021.
12. F. Naroll, R. Naroll, F. Howard, Position of women in childbirth. A study in data quality control.. 1961.
13. A. Vidich, G. Shapiro, A Comparison of Participant Observation and Survey Data.. 1955.
14. D. Jensen, T. Wilson, Data Quality Policies and Procedures: Proceedings. Washington, DC, USA, 1986.
15. C. Scannapieca, Data quality: Concepts, methodologies. New York: Springer-Verlag, 2006.
16. C. Batini, C. C., Methodologies for data quality assessment and improvement. In C. C. C. Batini. ACM Computing Surveys (CSUR), 2009.
17. S. Fatimah, P. H, Data Quality: A Survey of Data Quality Dimensions. IEEE, 2012.
18. Y. Man, L. W., A noval data quality controlling and assessing fuzzy association rules. In L. W. Y. Man, 2010.
19. Y. W. Wang, Anchoring data quality dimensions in ontological foundations. In Y. W. Wang. Communications of the ACM., 1996.
20. KQ. Wang, S. T., Analysis of data quality and information quality problems in digital manufacturing. In S. T. KQ. Wang, 2008.
21. M. Heravizadeh, J. M., Dimensions of business processes quality (QoBP). In J. M. M. Heravizadeh, 2009.
22. D. McGilvray, Executing data quality projects: Ten steps to quality data and trusted information. In D. McGilvray. Morgan Kaufmann, 2008.
23. Data.europa.eu, Data Quality Guidelines. In Data.europa.eu., 2021.
24. Tim Berners-Lee, site: https://5stardata.info/en/. 2001.
25. ISO/IEC, I. S. ISO/IEC 8601. 1988.
26. ISO/IEC, I. S., Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model. 2008.
27. ISO/IEC, I. S. . ISO/IEC 25024:2015 Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality. 2015.
28. M. Rodríguez, J. R. Oviedo, M. Piattini, Evaluation of Software Product Functional Suitability: A Case Study. 2016.

29. M. Dvaz-Ley, F. Garcva, M. Piattini, MIS-PyME software measurement capability maturity model - supporting the definition of software measurement programs and capability determination, Adv. Eng. Softw. 2010.
30. M. Hammer, J. Champy, Reengineering the Corporation: A Manifesto for Business Revolution, Harper Collins, 2001.
31. M. Stoica, N. Chawat, N. Shin, An investigation of the methodologies of business process reengineering. In Proceedings of Information Systems Education Conference, 2003.
32. T. Redman, Data Quality for the Information Age. Artech House, 1996.
33. L. English, Improving Data Warehouse and Business Information Quality. Wiley & Sons, 1999.
34. Liebchen & Shepperd, Data sets and data quality in software engineering. 2008.
35. D.C. Corrales, A. Ledezma, J.C. Corrales, A systematic review of data quality issues in knowledge discovery tasks. Revista Ingenierías Universidad de Medellín, 2016.
36. C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, in: ACM Computing Surveys. Vol. 41, No. 3, 2009.
37. F. Gualo, M. Rodríguez, J. Verdugo, I. Caballero, M. Piattini, 2021 Data quality certification using ISO/IEC 25012: Industrial experiences. J. Syst. Soft., 2021.
38. R. Wang, A product perspective on total data quality management. Comm. ACM 41, 2, 1998.
39. J. Oakland. Total Quality Management. Springer, 1989.
40. G. Shankaranarayan, R. Y. Wang, M. Ziad, Modeling the manufacture of an information product with IP-MAP. Boston: In Proceedings of the 6th International Conference on Information Quality (ICIQ 2000), 2000.
41. M. Jeusfeld, C. Quix, M. Jarke, Design and analysis of quality information for datawarehouses. In Proceedings of the 17th International Conference on Conceptual Modeling, 1998.
42. L. English, Improving Data Warehouse and Business Information Quality. Wiley & Sons, 1999.
43. Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang, AIMQ: A methodology for information quality assessment. Inform. Manage., 2002.
44. R. Wang, D. Strong, 1996. Beyond accuracy: What data quality means to data consumers. J. Manage. Inform. Syst. 12, 4, 1996.
45. L. Pipino, Y. Lee, R. Wang, Data quality assessment. Commun. ACM 45, 4, 2002.
46. M. Eppler, P. Munzenmaier, Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology. In Proceedings of the 7th International Conference on Information Systems (ICIQ), 2002.
47. Istat Guidelines for the data quality improvement of localization data in public administration (in Italian). www.istat.it, 2004.
48. P. Falorsi, S. Pallara, A. Pavone, A. Alessandroni, E. Massella, M. Scannapieco, Improving the quality of toponymic data in the italian public administration. In Proceedings of the ICDT Workshop on Data Quality in Cooperative Information Systems (DQCIS), 2003.
49. Y. Su, Z. Jin, A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. In Proceedings of the 9th International Conference on Information Quality (ICIQ), 2004.
50. D. Loshin, Enterprise Knowledge Management - The Data Quality Approach. Series in Data Management Systems, Morgan Kaufmann, 2004.
51. M. Scannapieco, A.Virgillito, M. Marchetti, M. Mecella, R. Baldoni, The DaQuinCIS architecture: a platform for exchanging and improving data quality in Cooperative Information Systems. Inform. Syst., 2004.
52. F. De Amicis, D. Barone, C. Batini, An analytical framework to analyze dependencies among data quality dimensions. In Proceedings of the 11th International Conference on Information Quality (ICIQ), 2006.

53. J. Long, C. Seko, A cyclic-hierarchical method for database data-quality evaluation and improvement. In Advances in Management Information Systems-Information Quality Monograph (AMISIQ) Monograph, R. Wang, E. Pierce, S. Madnick, and Fisher C.W, 2005.
54. C. Batini, M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques. Springer Verlag, 2006.
55. C. Batini, F. Aabitza, C. Cappiello, C. Francalanci, 2008. A comprehensive data quality methodology for Web and structured data. Int. J. Innov. Comput., 2008.
56. C. Batini, D. Barone, F. Cabitza, S. Grega, A Data Quality Methodology for Heterogeneous data. International Journal of Database Management Systems (IJDMS), Vol.3, No.1, 2011.
57. R.Y. Wang, A product perspective on total data quality management. Commun. ACM 41, 1998.
58. 58. A. Maydanchik, Data quality assessment. 2007.
59. I. Caballero, A. Caro, C. Calero, M. Piattini, IQM3: Information quality maturity Model. 2008.
60. S. Wang, H. Wang, Information quality chain analysis for total information quality Management. Int. J. Inform, 2008.
61. C. Guerra-Garcva, I. Caballero, M. Piattini, A survey on how to manage specific data quality requirements during information system development. Commun. Comput. Inform. Sci. 230 (Eval. Novel Approach. Softw. Eng.), 2011.
62. A. Nikiforova, Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. Baltic J. Mod. Comput., 2020.
63. ISO/IEC, ISO/IEC 25000 Software and System Engineering – Software Product Quality Requirements and Evaluation (SQuaRE) –Guide to SQuaRE, International Organization for Standarization. Geneva, Switzerland: 2005.
64. J. Akoka, L. Berti-Equille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoue-Thion, Z. Kedad, S. Nugier, V. Peralta, S. Sisaid-Cherfe, A framework for quality evaluation in data integration systems. In International Conference on Enterprise Information Systems, ICEIS, 2007.
65. B. Otto, A. Schmidt, Enterprise master data architecture: Design decisions and Options. Little Rock, USA: In International Conference on Information Quality, ICIQ, 2010.
66. B. Otto, Y. Lee, I. Caballero, Information and data quality in networked business. In Electronic Markets, 2011.
67. C. Guerra-Garcva, I. Caballero, L. Berti-Equille, M. Piattini, DAQ_UWE: A framework for designing data quality aware web applications. Adelaide, Australia: In International Conference of Information Quality, ICIQ'11, 2011.
68. C. Guerra-Garcva, I. Caballero, M. Piattini, Capturing data quality requirements for web applications by means of DQ_WebRE, Inform. Syst. Front, 2012.
69. C. Batini, M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques. Berlin: Springer-Verlag, 2006.
70. K.Y. Dahbi, H. Lamharhar, D. Chiadmi, Toward an evaluation model for open government data portals. in: International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning, Springer, Cham, 2018.
71. K. Kerr, T. Norris, The development of a health data quality programme. in Information Quality Management: Theory and Applications, IGI Global, 2007.
72. S. Van den Berghe, K. Van Gaeveren, Data quality assessment and improvement: A Vrije universiteit Brussel case study. Proc, 2017.
73. D. Kosar, Developing a Framework to Manage Data Quality in Healthcare Fourth International Conference on Information Quality. ICIQ'99, MIT, Cambridge, MA, USA, 1999.
74. B. Davidson, Y.W. Lee, R.Y. Wang, Developing data productions maps: Meeting patient discharge data submission requirements. Int. J. Healthcare Technol. Manag., 2004.

75. C. Bevan, D. Strother, Best Practices for Evaluating Method Validity, Data Quality and Study Reliability of Toxicity Studies for Chemical Hazard Risk Assessments. Washington (DC), USA: American Chemical Council, Centre for Advancing Risk Assessment Science and Policy, 2012.
76. T. Burzynski, Establishing the environment for implementation of a data quality management culture in the military health system. MIT, Cambridge, MA, USA: In Third International Conference on Information Quality, ICIQ'98, 1998.
77. M. Gendron, M.J. D'Onofrio, Formulation of a decision support model using quality attributes. MIT, Cambridge, MA, USA: In Seventh International Conference on Information Quality, ICIQ'02, 2002.
78. G. Shankaranarayan, M. Ziad, R.Y. Wang, Managing data quality in dynamic decision environments: An information product approach, J. Database Manag., 2003.
79. A. Caro, C. Calero, I. Caballero, M. Piattini, A proposal for a set of attributes relevant for web portal data quality. Softw. Qual. J., 2008.
80. M. Eppler, P. Muenzenmayer, Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In Proceeding of the Seventh International Conference on Information Quality, 2002.
81. P. Leonowich-Graham, M.J. Willshire, A data quality framework for small business. MIT, Cambridge, MA, USA: In Eighth International Conference on Information Quality, ICIQ'03, 2003.
82. M. Mecella, M.n. Scannapieco, A. Virgilito, R. Baldoni, T. Catarci, C. Batini, Managing data quality in cooperative information system. In CooPI/DOA/ODBASE. R. Meersman and Z. Tari, Springer-LNCS 2519, 2002.
83. D.M. Strong, Y.W. Lee, R.Y. Wang, Data quality in context. Commun. ACM 40, 1997.
84. T.C. Redman, Data Quality: The Field Guide. Digital Press, 2001.
85. U. Dama, Working group on data quality dimensions. In The Six Primary Dimensions for Data Quality Assessment: Defining Data Quality Dimensions, 2016.
86. ISO-25012, ISO/IEC 25012: Software engineering-software product quality requirements and evaluation (SQuaRE)-data quality model, 2008.
87. I. Jacobson, G. Booch, J. Rumbaugh, The Unified Software Development Process. Reading (MA) Addison-Wesley, 1999.
88. B. Nuseibeh, S. Easterbrook, Requirements engineering: A roadmap. Ireland: In Proceedings of the Conference on the Future of Software Engineering Limerick, ACM, 2000.
89. R. Pressman, Software Engineering: A Practitioner's Approach. 5/E, McGraw-Hill, 2001.
90. D. Lowe, J. Eklund, Client needs and the design process in web projects. J. Web Eng, 2002.
91. J. Nicolαs, A. Toval, On the generation of requirements specifications from software engineering models: A systematic literature review. Inf. Softw. Technol. 51, 2009.
92. J.C.S.d.P. Leite, P.A. Freeman, Requirements validation through viewpoint resolution. IEEE Trans. Softw. Eng. 17, 1991.
93. P. Darke, G. Shanks, Stakeholder viewpoints in requirements definition: A framework for understanding viewpoint development approaches. Requir. Eng., 1996.
94. G. Kotonya, I. Sommerville, Requirements engineering with viewpoints. Softw. Eng. J., 1996.
95. G. Kotonya, I. Sommerville, Requirements engineering. In Processes and Techniques, 2002.
96. G. Kotonya, Practical experience with viewpoint-oriented requirements Specification. Requir. Eng. S. Lond., 1999.
97. G. Kotonya, I. Sommerville, Requirements engineering with viewpoints. Softw. Eng. J., 1996.
98. C. Guerra-García, I. Caballero, M. Piattini, Capturing Data Quality Requirements for Web Applications by means of DQ_WebRE. Uppsala, Sweden: ACM, BEWEB, 2011.

99. Irfan Rafique et al., Information Quality Evaluation Framework: Extending ISO 25012 Data Quality Model. 2012.
100. M.A. Serhani, H.T. El Kassabi, I. Taleb, A. Nujum, Big Data quality evaluation across the Big Data value chain. 2016.

101. I. Taleb, H. T. El Kassabi, M. A. Serhani, R. Dssouli, C. Bouhaddioui, Big Data Quality: A Quality Dimensions Evaluation. Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, 2016.
102. I. Taleb, M. A. Serhani, Big Data Pre-Processing: Closing the Data Quality Enforcement Loop. IEEE 6th International Congress on Big Data, 2017.
103. D. Ardagna, C. Cappiello, W. Samá, M. Vitali, Context-aware data quality assessment for big data. Elsevier, 2018.
104. H. Tian, H. Wang, K. Zhou, Data Quality Assessment for On-line Monitoring and Measuring System of Power Quality Based on Big Data and Data Provenance Theory. 2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis, 2018.
105. M. Frye, R. Heinrich, Structured data preparation pipeline for machine learning-applications in production. 17th IMEKO TC 10 and EUROLAB Virtual Conference "Global Trends in Testing, Diagnostics & Inspection for 2030" 2020.
106. E. R. Bekar, P. Nyqvist, A. Skoogh, An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. Advances in Mechanical Engineering 2020, 2020.
107. Q. Chen, Y. Liou, S. Hou, F. Duan, Z. Cai, Data-driven methodology for state detection of gearbox in phm context. 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing), IEEE, 2021.
108. L. B. Iantovics, C. Enăchescu, Method for Data Quality Assessment of Synthetic Industrial Data. MDPI, Sensors, 2022.
109. T. Segreto, R. Teti, Data quality evaluation for smart multi-sensor process monitoring using data fusion and machine learning algorithms. Springer, 2022.
110. D. Xu, Z. Zhang, J. Shi, A Data Quality Assessment and Control Method in Multiple Products Manufacturing Process. 5th International Conference on Data Science and Information Technology, IEEE, 2022.
111. J.-P.Herrmann, S. Tackenberg, E. Padoano, J. Hartlief, J. Rautenstengel, C. Loeser, J. Bohme, An ERP Data Quality Assessment Framework for the Implementation of an APS system using Bayesian Networks. Elsevier, 2022.
112. J. Yang, G. Lan, Y. Li, Y. Gong, Z. Zhang, S. Ercisli, Data quality assessment and analysis for pest identification in smart Agriculture. Elsevier, 2022.
113. S. Sen, E. Husom, A. Goknil, D. Politaki, S. Tverdal, P. Nguyen, N. Jourdan, Virtual sensors for erroneous data repair in manufacturing a machine learning pipeline. Elsevier, 2023.
114. C. Guerra-García, A. Nikiforova, S. Jiménez, H. G. Perez-Gonzalez, M. Ramírez-Torres, L. Ontañon-García, ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design (DAQUAVORD), Elsevier, 2023.
115. H. Sun, F. Yang, P. Zhang, Y. Jiao, Y. Zhao, An Innovative Deep Architecture for Flight Safety Risk Assessment Based on Time Series Data. Computer Modeling in Engineering & Sciences, 2023.
116. Y. Yao, X. Zhang, W. Cui, A LOF-IDW based data cleaning method for quality assessment in intelligent compaction of soils. Elsevier, 2023.
117. M.A. Hamada, Machine Learning Model to Enhance the Quality of Software Development Risk Management. 2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE), 2024.

118. P. Sathiya, M. Amanullah, D. Manivannan, Enhancing SVM Classification of Meningitis with Feature-Adaptive Adagrad in CSF Analysis. 5th International Conference on Electronics and Sustainable Communication Systems (ICESC 2024) IEEE, 2024.
119. Q. Meng, Bayesian Optimization based Support Vector Machine for the Early Warming of Enterprise Financial Risk Analysis. 2024 International Conference on Data Science and Network Security (ICDSNS), 2024.
120. T. Zhang, R. Peng, M. Kadoch, Optimizing Data Quality in Deep Learning through Advanced Analytics. 2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) IEEE, 2024.
121. V. B. Vallevik, A. Babic, S. E. Marshall, S. Elvatun, H. M. B. Brøgger, S.Alagaratnam, B. Edwin, N. R. Veeraragavan, A. K. Befring, J. F. Nygård, Can I trust my fake data – A comprehensive quality assessment framework for synthetic tabular data in healthcare. Elsevier, 2024.
122. A. R. Iken, R. W. Poolman, M. G.J. Gademan, Data quality assessment of interventional trials in public trial Databases. Elsevier, 2024.
123. 123. J. Elsner, H. Brings, F. Sohnius, R. H. Schmitt, Data quality in environmental assessment methods - Implications for the operational management in manufacturing. Elsevier, 2024.
124. Y. Özçevik 2024, Data-oriented QMOOD model for quality assessment of multi-client software applications. Elsevier, 2024.
125. X.-H. Zhou, S.-L Shen, Assessment of living quality in Guangdong: A hybrid knowledge-based and data-driven approach. Elsevier, 2024.
126. C. Harris, L. Iannini, Web article The Top 10 Data Quality Tools. Expert Insights.com. 2024.
127. E. Stewart, Web article Top 10 Best Data Quality Tools for 2024, em360tech.com. 2024.
128. IBM Web site, https://www.ibm.com/products/infosphere-info-server-for-datamgmt. 2024.
129. SAP Master Data Governance Web site, https://www.sap.com/products/technology-platform/master-data-governance.html. 2024.
130. Talend Data Fabric Web site, https://www.talend.com/products/data-fabric/. 2024.
131. Ataccama Data Quality & Governance, https://www.ataccama.com/platform/data-quality. 2024.
132. Infromatica Cloud Data Quality Web site, https://www.informatica.com/products/data-quality/cloud-data-quality-radar.html. 2024.
133. Oracle Enterprise Data Quality Web site, https://www.oracle.com/middleware/technologies/enterprise-data-quality.html. 2024.
134. Melissa Unison Web site, https://www.melissa.com/customer-data-validation-platform. 2024.
135. Precisely Data Integrity Suite Web site, https://www.precisely.com/product/data-integrity/precisely-data-integrity-suite. 2024.
136. SAS Data Quality Web site, https://www.sas.com/en_gb/home.html. 2024.
137. Collibra Data Quality & Observability Web site, https://www.collibra.com/_us/_en/products/data-quality-and-observability. 2024.