Title: The changing scope of data quality and fit for purpose: evolution and adaptation of a CRIS solution

Type: Business/technical paper

Author: Thomas Gurney
Elsevier, Amsterdam, Netherlands

Keywords

Current Research Information Systems (CRIS), Research Information Management Systems (RIMS), Data quality, Fit for purpose, Data Deduplication, Open Access, Pure

Summary

This extended abstract aims to introduce and analyse implications and solutions required to improve data quality within the scope of *fit for purpose* in a Current Research Information System (CRIS) context. Drawing from, and building on, data and information quality foundations and descriptions, a data quality framework is introduced, and detailed product functionality is described. A discussion on the combination of framework and functionality highlights how data quality can be improved in a commercial CRIS product (Pure).

Data quality as a concept

Data quality in a CRIS context has been discussed and analysed at length within the CRIS community and associated literature. Formalised definitions of data quality in the CRIS context are still under development, with the greatest contributions coming from euroCRIS as a community, and in influential works by Stempfhuber (Stempfhuber, 2008) and Azeroual (Azeroual & Schöpfel, 2019b).

Central to data quality definitions proposed by these studies (and others contributing to this field) is the concept of *fit for purpose* and how CRISs must satisfy this as a key step to claiming to provide high quality data.

The Code of Good Practice (CGP), introduced by euroCRIS and CORDIS (euroCRIS & CORDIS, 1998) served as a guide for much of the research on data quality in CRISs, with the concept of *fit for purpose* a defining element. In general terms, analysis and discussion of the concept within a CRIS speaks to the ingestion, operationalisation, management and *use* of data - primarily from an institutional perspective. To be *fit for purpose* indicates that primary use cases, themselves dependent on the variety of institution and audience contexts, should be satisfied (Azeroual & Schöpfel, 2019a) and rely extensively on high quality data. As an example, all the use cases put forward in the GCP (included below) require that the (in)direct source data be of high quality:

- Identify potential funding
- Evaluate research funding
- Avoid duplication of research activity
- Establish priorities for research activities
- Source for scientific publishing
- Inform educational establishments, industry and the general public
- Facilitate information access and exchange

- Facilitate exploitation of research
- Internal administrative functions
- Analyse research trends within and across countries
- Compare institutions within countries in R&D
- Promote international cooperation
- Contacts and networks
- Locate new market for products
- Locate new products for market
- Locate persons/organisations with desired skills
- Administration tool for research projects

From a software product development and management perspective, based on the primary use cases presented above, understanding the 'purpose' in the concept *fit for purpose* is key to ensuring that the product (and by extension, data processed with in the product) provides the most value to that product's customers and users. CRIS providers facilitate this by investing significant resources to ensure that what data is provided and processed for its users is relevant and presented as early as possible whilst also being of high enough quality.

Linking the discussions between data quality in terms of being *fit for purpose* and how CRIS providers can add functionality to improve data quality, the over-arching question remains: what CRIS functionality is required to address the needs and recommendations of the CRIS community, and maintain general data quality principles at the same time? More specifically - with each institution operating within a unique context, how can a CRIS optimise itself to provide the most value? Furthermore, how can the dynamicism of data quality (Azeroual & Schöpfel, 2019b; Miller, 1996; Wang, 1996) be optimised for each customer and user type? - and, given the relative heterogeneity of the CRIS market, and development of community and national or regional CRISs, how can data quality be ensured in interoperability and customer and researcher migration between CRISs (van den Berghe & van Gaeveren, 2017)?

Given the range of questions around data quality in CRISs, with only a few presented above, and the necessarily short format of this paper, the following sections of this paper will address how data quality is defined within Pure, a commercial CRIS, and includes two specific functionalities that were designed with the goal of improving data quality on a continuous level.

Pure and data quality

Pure introduced the data quality framework, 'The 5Cs of data quality', as an internal (and external) reference as to how data can and should be optimised for each customer and user. The 5Cs framework aims to ensure that data within a Pure is: Complete, Correct, Connected , Current and Compliant. These concepts map clearly to common (e.g. Azeroual & Schöpfel, 2019a; Lee et al., 2002; Miller, 1996) contexts and definition descriptors of information quality which include:

- Intrinsic e.g. accuracy, consistency/reliability
- Contextual e.g. completeness, relevance, timeliness, currency
- Representational e.g. interpretability, compatibility
- Accessibility e.g. accessibilty, system availability, ease of use, security

The 5Cs framework is primarily based on intrinsic and contextual information quality, with representational information quality addressed through Pure's adoption and support of common standards and vocabularies including CERIF, Dublin Core and DataCite Metadata Schema. Accessibility information quality is addressed through Pure's user interface and layout, and user/role access support. The concept of 'Connected' is a specialised descriptor in the 5Cs framework. This is a trans-contextual term and it is specifically defined to address new opportunities and purposes.

Like most CRISs, Pure uses a relational database with defined schemas to store data. The core module in Pure supports content creation and interaction via a user interface, with modules for reporting, showcasing, and specialised content management. Pure supports over 30 different content types, accommodating an extremely wide variety of research activities, processes and outputs, all with varying natures, forms and sources. Underlying this are strict data models that govern how structured (e.g. customer-defined classifications) and unstructured data (e.g. title and abstract), can be added and modified – whilst accommodating and controlling for common standards and vocabularies.

Content can be added in multiple ways including manual creation, import from over 45 data sources or synced in from external institution-specific data sources or even other Pure instances. Access to Pure is managed across a broad variety of roles, with varying limitations on configuration of create, read, update and delete actions on a per-content-type level.

Given the variety of user roles, import sources and content types within Pure, improving data quality through the 5Cs framework occurs at all functional and development levels. The 5Cs primarily focus on error prevention and correction, such as within the interface, database schemas and underlying data models. Expanding the purpose of Pure, the 5Cs also cover data enrichment, notably Pure's integrations with external data sources and data transformation layers between these sources and the underlying data models in Pure.

Opportunities for improving data quality and scope of fit for purpose

Presented below are two contexts and scenarios where Pure has developed functionality to address specific data quality issues.

Improving data quality via compliancy recommendations

Pure provides support and guidelines for Open Access (OA) publishing of content in multiple scenarios. However, OA guidelines are extensive and will vary by institution, journal and version of document being added to a record – leading to user confusion when adding content to Pure and potentially reducing data quality of the system. Pure has addressed these correctness and compliance data quality issues in several ways. Firstly, the underlying data model that specifies metadata and statuses on content, adheres to recommended standards, and is designed to accurately reflect the OA information provided by data sources, whilst supporting the OA reporting and showcasing needs of the institution. Secondly, for records that are missing required information or files, Pure has implemented specific functionality to notify users of missing information.

Thirdly, Pure surfaces institution- and journal-specific relevant OA information at three major, configurable, touch points:

- The journal itself, where institution-specific OA policies and recommendations can be listed alongside journal-specific OA requirements (sourced via API from SherpaRomeo[1]).
- The main metadata page of the research output editor itself, wherein the same institution and journal information is prominently displayed.
- The file upload and OA status settings on a record, which also includes the institutional and journal recommendations. What makes this functionality stand out is that the information is dynamically presented and displays journal requirements for the specific file version that a user is trying to upload. The result being that the complexity of OA requirements is reduced to the specific task at hand and the requirements for that task.

These first two touch points could be considered standard for most CRISs but the third is unique as it reduces the confusion that a user must face in an already confusing context.

Improving data quality through independent data source updates

For all CRISs, data quality can be limited by the nature and coverage of data available for import from data sources. An institution is limited by the selection of data sources they can access, with each data source having differing levels of quality. Some data sources benefit from earlier indexing which allows the institution to add content as early as possible, whilst other import sources provide more complete and enriched data for the same record, albeit available later. Data quality in a CRIS is further complicated by the need for imported content to be mapped to the CRISs internal data model for all entity (or supplementary) information in that import.[2] For example, importing a journal article is not limited to creating a record of the article itself in the CRIS, but also involves mapping any author and affiliation data of that article to any internal representations of those entities. Additionally, content within a data source may itself be subject to updates and mapping issues from any upstream contributing data sources.

Specifically addressing data quality completeness, correctness, and currency, Pure allows for the updating and enrichment of records, independent of the original data source, with configurable trust and specified actions on a per field level. As an example, a normal user and record flow might include the following: a user can import a journal article from PubMed as it was indexed earlier than in Scopus or Web of Science. Any entity mapping to internal representations is completed and the record is saved. Sometime later, that same record is brought to the attention of a user with elevated roles in Pure. They are notified that the record has an updated publication status and publication date, with any statuses and dates provided not by PubMed, but by Scopus and Web of Science. Pure also informs the user that

[1] Sherpa Romeo (SR) is a free service by Jisc (https://v2.sherpa.ac.uk/romeo/) that provides community-sourced publisher-specific Open Access recommendations.

[2] A data quality survey sent to Pure customers addressed, amongst other data quality considerations, the level of quality of content created or imported through specific methods. Content imported and saved by Pure administrators was judged to result in a much higher quality record than the same content imported and saved by a researcher. A commonly cited reason for the difference in quality is that researchers tended to spend less (if any) time checking the internal mapping where identifiers were not found in the record. Survey sent October 2021; 84 respondents from 11 countries.

the abstract in the original record was incomplete and Web of Science has an updated version they could use instead. There are also additional data source-specific keywords, and the user is given the option to merge all keywords from the data sources or overwrite the original with a specific source's keywords. If a Pure administrator prefers specific data sources, Pure can be configured to automatically accept and overwrite fields from specific sources or manually check enrichment opportunities on a record with data from multiple sources.

This functionality provides the institution with the ability to capture and use content as early as possible, whilst continually updating and enriching from preferred data sources on a per-field level. The data quality of a record is also improved through corroboration across multiple sources but also provides for discrepancy correction between sources.

Conclusions

Data quality in CRISs can, and should, be improved. Essential to this is an understanding by CRIS providers that efforts to improve data quality revolve around the concept of *fit for purpose*.

However, given the evolution and expansion of activities, outcomes and processes of research, a CRIS' scope of purposes has expanded correspondingly. For data quality purposes, this has been complicated by the scope and nature of how data is generated, ingested, operationalised and managed within a CRIS. As such, a CRIS needs to account for the changing nature of research and evolve its purpose and data quality management functionalities accordingly.

Pure created the 5Cs framework with this evolution in mind. The 5Cs of data quality framework is derived from work undertaken in the information quality and CRIS fields and transposed for use in Pure specifically. The ongoing research, standard-setting and exceptional knowledge-building in the CRIS communities have direct influence on efforts to improve data quality, and Pure contributes to this from a product perspective.

Azeroual, O., & Schöpfel, J. (2019a). *Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries*. https://doi.org/10.3390/publications7010014

Azeroual, O., & Schöpfel, J. (2019b). Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries. *Publications 2019, Vol. 7, Page 14*, *7*(1), 14. https://doi.org/10.3390/PUBLICATIONS7010014

euroCRIS, & CORDIS. (1998). *euroCRIS Code of Good Practive (CGP) 1998 v3*. http://arge.tuwien.ac.at/arge/europe/Cgprpt7.doc

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management*, *40*(2), 133–146. https://doi.org/10.1016/S0378-7206(02)00043-5

Miller, H. (1996). The multiple dimensions of information quality. *Information Systems Management*, *13*(2), 79–82. https://doi.org/10.1080/10580539608906992

Stempfhuber, M. (2008). Information quality in the context of CRIS and CERIF. *Proceedings of the 9th International Conference on Current Research Information Systems*. https://dspacecris.eurocris.org/handle/11366/331

van den Berghe, S., & van Gaeveren, K. (2017). Data Quality Assessment and Improvement: A Vrije Universiteit Brussel Case Study. *Procedia Computer Science*, *106*, 32–38. https://doi.org/10.1016/J.PROCS.2017.03.006

Wang, R. Y. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–34. https://doi.org/10.1080/07421222.1996.11518099