Keep the Best – Forget the Rest? Towards Models for CRISs Integrating Heterogeneous Information

Maximilian Stempfhuber
Social Science Information Centre, Bonn

Summary

Standards have so far been the primary means to ensure a minimum level of quality and common metadata structures across current research information systems (CRISs) and to support data exchange. The users' perspective is currently covered by suggestions for standard features a CRIS should offer at the user interface level, like free text search. But is this a sufficient precondition or a viable way for building information services that suit the whole targeted audience without neglecting special requirements of individuals?

In this paper we argue that in situations with many different data providers and users of data, standards per se are no sufficient means of ensuring usefulness, acceptance and sustainability of a service. We believe that only explicitly modelling the users' and data providers' perspectives of CRISs at a much higher level, and accounting right from the start for differences that can not be standardized will lead to a common basis for communication between interest groups and to systems that deliver a common user experience.

1 Introduction

Now, that the Internet is present in many peoples' working and private life, seeking for information on the Web is for most of them a very natural thing. They are used to have access to all the information at any place and at any time by using search engines like Google[1] or Altavista[2] or web site directories like Yahoo[3]. Even for special interest groups, dedicated information services are available. In the area of scientific information, the offerings of Scirus[4], LLEK[5], Online JOurnals Search Engine[6] (OJOSE), CiteSeer.IST[7] or the new German scientific gateway vascoda[8] – to name just a few – play an important role in limiting the information space of the Internet, potentially increasing the relevance of the information found and reducing the amount of irrelevant material.

---

[1] http://www.google.com

[2] http://www.altavista.com

[3] http://www.yahoo.com

[4] http://www.scirus.com

[5] http://www.scientific-search-engines.com/

[6] http://www.ojose.com/

[7] http://citeseer.ist.psu.edu/

[8] http://www.vsacoda.de

It can not be neglected that the advancements in search engines – both technically and from an information retrieval perspective – together with the willingness of the researchers to publish their results on the Web have improved the overall situation of everyone who either seeks information or tries to make it publicly available. One important issue which supported this development was the availability of open, non-proprietary standards like the Hypertext Markup Language (HTML) for encoding information, the Hypertext Transfer Protocol (HTTP) for exchanging information, and of free software (e. g. Netscape or Mozilla web browsers) for producing and using information complying with these standards. By using these standards, which are hard to avoid when interacting with the Internet, everyone should be able to give relevant and valuable input to the common information pool, and everyone should be able to make use of it. But is this the case?

Newer results from end user studies (Binder et al. 2001, Boekhorst et al. 2003, IMAC 2002, Stahl et al. 1998, Stahl et al. 2002) show that this might not be the case. Internet search engines – while often used as a starting point – do not deliver the quality and depth of information often looked for, but help the user to find (some) relevant pieces of information quickly. It is the "now-or-never" mindset which causes researchers to only browse the first few results of a query or take only the results into consideration which are directly linked to (for them) freely available full text documents. Still most scientific information systems contain only metadata referring to the original work and do not allow instant access to the original materials. And they are distributed, heterogeneous and most of the time inaccessible to search engines. But users seem to have clear expectations to scientific information systems:

- Domain-specific portals covering a single discipline.
- Direct access to full text documents and other original materials.
- Clusters of portals adequate for interdisciplinary information needs.
- Not only databases and library catalogues, but "everything" relevant.
- Not the "noise" delivered by internet search engines.
- Integrated access to all information services from their workplace.
- Ways for informal communication with colleagues.

In the area of digital libraries, current research information systems (CRIS) and portals to scientific information, standardization is the primary means to achieve many of the goals mentioned above. This includes standards for raw data, descriptive metadata, quality criteria, and for communication and data exchange.

From the number of existing or emerging standards and the years and manpower spent in defining and implementing them, one could expect that in the near future a situation will be achieved in which every information provider plays by the standards and that homogeneous information services can be linked together by the end user, according to the publishing paradigm on the Internet, where everything is interlinked and can be combined to some new "publication" at a higher level. We argue that this might not be the case and that standards by themselves are no sufficient means for building the information systems the users demand. Our intention is to put back the user into the centre of attention and switch from a viewpoint of "What are our services? How could a standard to ensure quality look like? How do we convince other service providers it's the best?" to "What does the user require? Can – and if so – how can we meet his expectations? Where and how far will standards help? How do we deal with the rest?"

2 The challenge of heterogeneity

Heterogeneity can nowadays be seen as one of the central problems – or challenges – when building integrated information systems. It arises for example from differences in data structures and content analysis between data collections as soon as these collections have to be integrated within one information system. At the user's side, the differences sometimes can be hidden by using simple search functions (e.g. the famous "Google-like" search) and full text indexing. But as soon as more details of the underlying structures have to be exposed (e.g. keyword search with thesaurus support) the user is faced with fields which are not valid for all databases or with only a basic index instead of a thesaurus, because not all databases are indexed with a thesaurus and those who are use different ones. Standardization might seem as a natural way of dealing with this heterogeneity, but the data side of an information system is only one level on which it occurs. Heterogeneity occurs at the user level, at the level of domains and disciplines, in the data and between data providers. And to make it even more difficult, heterogeneity changes over times as the relevant entities at each level do change by themselves.

2.1 Heterogeneity at the user level

An information system, like any other multi-purpose system, will be used by many different users with many different information needs and expectations to the services the system should provide. For scientific information many different groups of users come to mind, e.g. students, teachers, researchers, journalists and the media in general, companies, funding agencies, and policy makers (Boekhorst et al. 2003, CGP 1998, IMAC 2002, Koopmans 2002). They all require specific information like primary data (e.g. literature references, project descriptions and information about persons), aggregate data (e.g. number of projects or budget spent in an area, number of publications or patents) or even semantic relationships within the data (e.g. experts in a field or cooperation networks across countries).

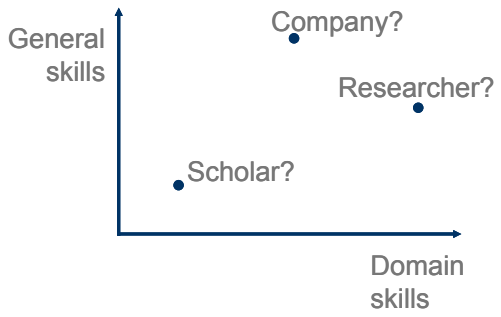

Figure 1: Two dimensions of a user's skills

Regarding users' skills for finding information, the two dimensions of general skills (e.g. computer literacy or information seeking strategies) and domain skills (e.g. expertise in their field) can be distinguished. Figure 1 shows exemplarily three types of users rated in regard to general and domain skills. An information system would have to support these users with search features of

different complexity (simple vs. advanced search) and alternative knowledge structures from the domain (e.g. thesaurus, classification, chemical formulas, maps) to be adequate for all users.

Finally, users themselves and the behaviour they exhibit change over time. This may include the user's transition from novice to domain expert during education (domain skills) or from casual user to power user (general skills) as he gets used to the information system over a longer period of regular use. In the same way, search strategies and information needs will change as the user learns about database structure and content and as he has to deal with problems of different complexity and time restrictions for information procurement. The communication with other users will also influence what information will be requested, how it will be accessed and used.

2.2 Heterogeneity at the domain level

In the same way users differ, domains or disciplines do regarding their culture and sociology. The process of generating knowledge and producing scientific results may be as different between two disciplines as the use of information technology. While in some scientific communities knowledge will be shared very early and research is carried out in teams or networks, other disciplines follow a more "closed shop" approach where publication of results happens at a very late stage. Similar varying observations can be made between engineering and social sciences / humanities concerning the search for relevant literature. While in the former often quick results are needed that can directly be applied to some problem, the later try to get a very detailed and complete overview of a topic, not to miss something important.

Last but not least, the amount of research activity and the quantity of output is different between domains, much like the recognition of these results by the public.

2.3 Heterogeneity at the data level

We have already indicated that the relevant user groups may have different expectations of what a current research information system should contain or deliver. Looking at the data side of a CRIS, things like reference databases for literature, full text documents with online ordering, project information, raw data (e.g. surveys, time series data or specimen), people, teams, institutes and networks come to mind. While it is clear that project information and survey data have their own structures and means for content analysis, even documents – in the broadest possible sense – of the same type can be very different. This is especially the case where collections of data already exist and have to be integrated with other collection. Features like languages, (meta)data structures, content analysis and indexing, storage systems, retrieval models, and the aspects of cost and accessibility (free vs. liable for costs / public vs. scientific use of survey data) play an important role here. Also within one collection heterogeneity grows over time when new metadata elements are introduced or the thesaurus is modified, because older records are rarely touched again and completed with the now missing information.

2.4 Heterogeneity at the data provider level

Data providers have much influence on the heterogeneity of data, because their own information needs, purposes for collecting data and available resources directly influence what will be col-

lected, how it will be structured, indexed and stored, and who will get access to the data under which circumstances and conditions. Free or low-priced technologies, like databases management systems, web servers and programming environments enable many interest groups to be information providers: Authors at their homepages or with self-archived open-access publications, researchers with materials collected during research for their own or their colleagues' use, research institutes, information centres and libraries, funding agencies and decision makers at the local, national or EU level.

The purpose for collecting the data and building a CRIS may be as individual as the data they collect. This includes recognition by the community, administration of current research, reports to boards, proof of excellence, services to the community or paying customers, and strategic planning. Some of these goals lead to overlapping requirements to what is collected and how it is organized and documented – which are good candidates for standardization. But the availability of local resources and the implications of publishing certain data may give rise to the decision to collect, organize, document or publish in some way distinct from other providers of CRISs. This directly influences their willingness to document certain research activities and results, to provide metadata and/or raw data, to provide direct access to the system and data, to deliver data on different media and in different formats, to make it freely available or charge costs, and to cooperate with third parties.

The question now is how we can handle this heterogeneity, which exists at all levels of an information system? One answer to this question is standardization. We argue that standardization by itself is a good strategy that will help to make certain aspects of information systems more homogeneous. But we do not believe that there is a chance to deal with all the problems in an adequate timeframe using only standardization. The next section will explain some of the obstacles with standardization and give rise to the point that we should start looking at the problem from a different perspective.

3 The problems with standards

The large number of initiatives concerned with structuring and standardizing information indicate, that more agreements besides the existing ones might be necessary to come to a situation where information from multiple providers can be automatically and intelligently handled by computing systems in the way it is expected by the information seeking user. Standardization currently covers the whole range from raw data to the end-user experience, e.g.:

- Raw data: OpenOffice[9] XML format (Eisenberg 2004) or DocBook[10] (Walsh & Muellner 2003) for publications, Portable Network Graphics[11] (PNG) for raster images, Scalable Vector Graphics[12] (SVG) for two-dimensional graphics, Web Ontology Language[13] (OWL) for semantic relationships.

[9] http://www.openoffice.org/

[10] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=docbook

[11] http://www.w3.org/Graphics/PNG/

[12] http://www.w3.org/Graphics/SVG/

- Metadata: Data Documentation Initiative[14] (DDI) for surveys or censuses in the social sciences, Dublin Core Metadata Initiative[15] (DC) often used for – but not restricted to – publications, Common European Research Information Format[16] (CERIF) for e.g. research projects, persons, publications.
- Communication: Web Services Activities[17] for higher level communication between applications, Open Archives Initiative[18] Metadata Harvesting Protocol (OAI-MPH) for collecting metadata.
- Information systems as a whole: Code of Good Practice (CGP 1998).
- End user experience: Yale Web Style Guide[19] for browser-based user interfaces, ISO 9241 covering ergonomic requirements for office work with visual display terminals.

But will it be a successful strategy to try to standardize all levels of an information system? How many standards will there be needed to cover every aspect which is relevant for more than one user, data type or information provider? How long will it take? How many competing standards will have to proof their advantages above others? Should we wait for worldwide standards or keep working at the European level? Can we enforce or only promote our standards?

There are many reasons why existing standards still are not used throughout our community – and maybe never will be. The following examples should help illustrate some of the problems with standards – not the ones concerning the standard per se, but those who influence the process of standardization and of adapting standards and which might have much more implications on the process as a whole.

3.1 Are standards feasible?

The first question to ask is: Are standards for every part of the information systems we'd like to build really feasible in a given timeframe? Will there be:

- A single format for every given data type?
- A single metadata schema adequate for every data type?
- A single indexing language for all domains?
- A single retrieval model and ranking algorithm?
- A single architecture for information systems?
- A single communication protocol?
- A single authentication authority?
- A single …?

---

[13] http://www.w3.org/2001/sw/WebOnt/

[14] http://www.icpsr.umich.edu/DDI/

[15] http://www.dublincore.org/

[16] http://www.eurocris.org/

[17] http://www.w3.org/2002/ws/

[18] http://openarchives.org

[19] http://www.webstyleguide.com/

As soon as alternatives are accepted, we may introduce the problem of heterogeneity again. While, for example, Dublin Core gains more and more acceptance, the number of special application profiles grows. This again introduces the problem of interpreting the semantics of metadata elements between applications. What remains might be only the syntactic level of the standard, a common way of structuring – but not interpreting – metadata.

3.2 Where are the problems with standards?

It seems there are two different attitudes to standards, mostly depending on what is standardized, how expensive it is to implement the standard and what the implications for the future are, i.e. if the changes made to implement the standard can be reversed with minimal effort and cost. While the technology-oriented standards, like XML or WebServices, are adopted very fast – at least for testing purposes – and picked up even at very early stages in the standardization process (beta versions), standards concerning (meta)data structures normally take much longer to be adopted. One important reason surely is the amount of resources needed to restructure data and at the same time keep its integrity and consistency – if the technology used allows restructuring the data at all. The same holds true for standards concerning indexing and classification. And finally, many data providers are involved in some kind of competition (e.g. for revenue, funding or usage), which forces them to keep a competing edge over the other competitors, something which could be lost during standardization.

One further aspect is, that with the complexity of the domain the standard also gets more and more complex and contradictions may be introduced or the correct implementation of the standard might not be guaranteed. Obviously this is the case with standards for user interface design, but it also can be seen in other domains.

3.3 Where do standards help?

Until now, only the negative aspects of or the obstacles experienced with standardization have been mentioned. But surely there are great benefits to be earned in situations where standards are accepted – and adhered to – by most parts of the community. This may include standards for communication, data exchange, universal access, common user experience or every situation where new work is being started and tried and tested standards are available. Following the standard will be the best solution in many cases.

But what should we do with the rest? Is there a way to deal with the remaining heterogeneity? Is the problem with heterogeneity also experienced in other domains? And what are their solutions?

4 Why we need a model

The problems with standardization mentioned before are not specific to information systems or the CRIS community. In many areas where standardization is important for reasons of security, interchange ability, and economic or technologic advancement, it is also realized that alternatives are needed to make the necessary progress. Here, the problems – and costs – of global standardi-

zation exceed by large the efforts involved in CRISs and the lengthy process of standardization may itself be a main obstacle or may not be able to keep up with actual developments.

The solution may be to open the process of standardization to the idea that there always will be details which can not be standardized and that this fact should be kept in mind and accounted for right from the beginning of the standardization process. The German standardization body, DIN[20], adapted this strategy in its recent position papers (DIN 2003a, DIN 2003b) about standardization:

> "The classic approach to standardization, to achieve compatibility and interoperability by technical uniformity, reaches its limits where regional or industry-specific solutions have been implemented with great effort (e.g. infrastructures) and subsequently global interoperability has to be assured as a result of globalization or a changed business situation.
>
> SICT recommends regarding standardization also from the viewpoint of providing 'interoperability for existing heterogeneity'.
>
> This task consists of finding a reasonable balance between the desired scope of standardization and the remaining heterogeneity treatment. The costs and the quality losses possibly resulting from the heterogeneity treatment should be seen in relation to the expenditure and the chances of success of further intensified standardization." (DIN 2003a: 7)

But how could such a strategy be implemented into the process of building future CRISs? Our suggestion is to make much more use of the methods developed for user centred design when building CRISs as it has been visible in the past. We feel that currently high-level models for important aspects of information systems are missing, which would give a much better basis for involving the user side and matching the interests of all participating parties according to common goals instead of existing differences. Which models would be needed? Models for:

- the user experience,
- dealing with structural heterogeneity,
- dealing with semantic heterogeneity,
- connecting relevant objects at the semantic level,
- aggregating distributed data to information, and for
- making cooperation possible.

The advantage of models is that they can – in contrast to standards – be communicated in a way that every party is able to instantly connect parts of the model with their personal knowledge and experiences, interpret the model and transfer it to their own application scenarios. For a user scenario this would for example mean that typical information needs are described together with the relevant data and ways for processing, displaying or visualizing the data. Using this framework, the need for homogeneous information can be identified in a CRIS, standards can be defined and concepts and processes for integrating non-standardized data can be developed in a way that they are consistent with the model. New requirements or data types would always first be matched against the model and then integrated if all possible contradictions or inconsistencies are resolved.

The next task of the CRIS community in our opinion would be to support the dialog between all interest groups involved and to coordinate the development of an explicit model of the CRIS of tomorrow – based on standards where possible and anticipating differences where necessary.

---

[20] http://www.sict.din.de

5 Conclusion

In this paper we tried to motivate some alternative view on building information systems for a broad audience and with heterogeneous content, which is the "model perspective" rather than the "standard perspective". We stated that standardization is very important at different levels of such an information system, but it is no sufficient prerequisite for building useful and user friendly systems. As in the user centred approach for interactive systems design, which tries to ensure that the requirements of the user are met during the whole development process, the attention when building a CRIS should be given to the targeted user groups and their respective information needs, before any bibliographical or technical standards are developed, chosen or implemented.

Standards are often the formal expressions of models and ideas, but very rarely they communicate the "whole" idea explicitly enough so that two adaptors of the standard would use it the same way. To solve this problem, we suggest building models of our users, their information needs, and the context – the use cases – in which they access our information systems. Discussing information needs with peer groups, taking every relevant type of information into account and trying to sketch the "optimal" information system could lead to models which would allow to select the proper (existing) standards, spot areas where new standards would be helpful and plan for alternatives where standardization seems not feasible.

6 References

Binder, G.; Klein, M.; Porst, R.; Stahl, M. (2001): *GESIS-Potentialanalyse: IZ, ZA, ZUMA im Urteil von Soziologieprofessorinnen und –professoren*. GESIS-Arbeitsbericht, Nr. 2. Bonn, Köln, Mannheim.

Boekhorst, P. te; Kayß, M. & Poll, R. (2003): *Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung. Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft*. Universitäts- und Landesbibliothek Münster und infas Institut für angewandte Sozialwissenschaft GmbH, www.dfg.de/forschungsfoerderung/wissenschaftliche-_infrastruktur/lis/aktuelles/download/ssg_bericht_teil_1.pdf

CGP (1998): *Code of Good Practice for Current Research Information Systems*. EuroCRIS, Version 3.0, January 1998, http://www.eurocris.org/products/codegpr.doc

DIN (2003a): *Standardization in Information and Communication Technology (ICT). German Positions*. DIN Deutsches Institut für Normung e.V. Strategy Committee on Standardization in Information and Communication Technology (SICT), http://www.ni.din.de/sixcms_upload/media/1436/sict_artikel_engl.pdf

DIN (2003b): *Knappe Ressourcen effektiv nutzen*. Strategieausschuss für die Standardisierung in der Informations- und Kommunikationstechnik (SICT) im DIN Deutsches Institut für Normung e. V. http://www.sict.din.de/sixcms_upload/media/1437/sict_kb_strategieplan.pdf

Eisenberg, D. J. (2004): *OpenOffice.org XML Essentials – Using OpenOffice.org' s XML Data Format*. Preliminary version, http://books.evc-cit.info/

IMAC (2002): *Projekt Volltextdienst. Zur Entwicklung eines Marketingkonzepts für den Aufbau eines Volltextdienstes im IV-BSP*. IMAC Information & Management Consult-ing. Konstanz. September 2002. Management summary.

Koopmans, N. I. (2002): *What's your question? The need for research information from the perspective of different user groups*. In: Adamczak, W.; Nase, A. (eds.): Gaining Insight from Research Information: Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel, August 29-31, 2002. Kassel: Kassel University Press, pp. 183-192, http://www.uni-kassel.de/CRIS2002/files/pdf/Koopmans.pdf

Stahl, M.; Binder, G.; Cosler, D.; unter Beratung von Dr. Joachim Scharioth (1998): *TRI:M-Studie zur Kundenzufriedenheit (Mehrfachkunden) 1997*. Bonn, IZ Sozialwissenschaften (IZ-Arbeitsbericht; Nr. 13) http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab13.pdf

Stahl, M.; Binder, G.; Marx, J. (2002): *Das Informationszentrum Sozialwissenschaften im Urteil von Soziologieprofessorinnen und -professoren aus Deutschland, Österreich und der Schweiz*. Bonn, IZ Sozialwissenschaften (IZ-Arbeitsbericht; Nr. 25) http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_25.pdf

Walsh, N; Muellner, L. (2003): *DocBook: The Definitive Guide*. Updated: Wed, 31 Dec 2003, http://www.docbook.org/tdg/en/html/docbook.html

7 Contact Information

Dr. Maximilian Stempfhuber
Social Science Information Centre (IZ)
Lennéstr. 30
53113 Bonn, Germany

e-mail: stempfhuber@iz-soz.de
Homepage: http://www.gesis.org/IZ/stempfhuber/