DAM-LR as a Language Archive Federation: strategies and prospects

David Nathan and Remco van Veenendaal

SOAS, London; INL, Leiden
djn@soas.ac.uk

1. Introduction

The DAM-LR partners are well on the way to forming a federation. The term 'federation' has at least two quite different meanings and it is important not only to distinguish them but also to put our own stamp on what kind of federation we create.

'Federation' has a specialised meaning in information technology, referring to bringing information resources together via information management and networking techniques. It also has an organisational meaning, referring to agencies and their aims and strategies for collaboratively dealing with identities, resources, and services. In this paper, we refer to 'federation' in the first domain as federationT ("federation technologies") and in the second domain as federationA ("federation agency/-ies").

2. FederationT: a background

Federations in the IT sense go back to the earliest days of electronic networking. For example, in 1967 the Online Computer Library Center (OCLC, http://www.oclc.org/) at Ohio State University started sharing bibliographic authority files with fellow libraries, and has long been involved with the issues that still face us now: standards, metadata, quality levels, technology, membership and collaboration. The OCLC now has 9,000 members who share 65 million records to assist in their librarianship work. By the mid 1990s the term 'federated database' was well known. Dempsey et al, for example, describe a "federating solution … [that] allows services to develop autonomously while projecting a single unified image to the user". The motivation for federating resources is to provide value to users by providing a larger metadata set with a correspondingly greater ability to "relieve … potential users of having to have full advance knowledge" of the existence or nature of resources (Dempsey et al). According to this definition, search engines such as Google also, in a weak sense, federate all the web pages that they index.

Lynch (1998) refers to Dublin Core (DC) – also with its roots in libraries – as a tool for federating existing resources: "networked information discovery and retrieval [through] federating disparate, independently maintained databases … [via] a common semantic view of the various databases involved". DC was intended to enhance resource discovery in an open networked world, i.e. in a world without librarian operated-catalogues where quality and consistency are principal values and practices. Dublin Core provides, then, a low-density, lowest-common-denominator but unified method for description and discovery in a unified search domain (the www) by a wide variety of professionals, data-creators and ordinary users. To achieve this, the DC consortium dealt with issues such as (i) syntactic standards e.g. for data and metadata encoding (ii) semantics, e.g. ontologies, semantic web etc. (iii) strategic goals, e.g. selection of the lowest-common-denominator approach to lower the costs and other barriers to coding.

Note that computing power here is a catalyst rather than a central factor; most of the activity is done by humans. FederationT in the sense discussed here contrasts with its use elsewhere to refer to linking networks or grids of computers in order to provide a scaling up of computational power. Here, we seek to scale up resource discovery, retrieval, and preservation, rather than processing.

More recently, parts of the linguistics community have been working in similar areas – OLAC, which was similarly centred on strategic goals for resource discovery, and GOLD ontology, which focussed on mapping out the concept territory of linguistics, to enable linguists to cross-map their varied terminologies (i.e. to bridge between author-created metadata and unified metadata formalised by a body of professionals). OLAC has been moderately successful, although more in terms of raising awareness about issues in language data handling than in unifying resource discovery across language data repositories, possibly because of its broad but ambiguous ambit ranging from endangered languages to multimedia to any language data. GOLD has been motivated by the putative needs of the "endangered languages community" (http://emeld.org/workshop/2003/paper-terry.html), but has mainly drawn interest from typologists and computationalists.

Ultimately, resource discovery has not, at least so far, been a foreground problem for most linguists. In other areas, web search engines have provided alternative solutions, and various areas of industry and commerce have been unobtrusively implementing EDI systems.

A conclusion one might warily draw is that the linguistic community has not (at least yet) found a clear need for such resource discovery and ultimately

federalism among repositories. On the other hand, however, linguists will benefit from previous and current work when the day comes that they do find such needs. Progress is likely to be sudden rather than evolutionary, when, at some point, linguists find that not only their tools (email, word processors, databases) but also their modes of expression are electronic (most likely this will occur among the forthcoming generation that will have been fully imbued with electronic communications of all kinds). Once enough linguists' decide to disseminate their own resources via electronic repositories, then federated electronic repositories will become a major locus for searching for other linguistic materials.

3. Opportunities

The current environment for language and technology and the nature of the DAM-LR partners suggest a number of opportunities that can guide strategy for collaboration. Our archives have relatively clear conception of our aims, holdings, and audiences, enabling us to exploit the valuable insights from specific linguistic (and related) subdomains, such as specialised corpuses, endangered languages, sign languages, the collection and implementation of protocol, new genres of data and presentation, new modes of access, and recognition of the new client groups for whom language data is crucially important.

Federating offers us important opportunities, because our repositories hold data that is typically fragmented, not published (or not conventionally publishable), and rare (in fact, it is the fragmented, data-oriented nature of our materials that unifies them as much as the fact that they are linguistic resources). Federation will provide increased dissemination opportunities and therefore add value to our individual collections.

In addition, we have a focal client group, depositors, to whom we need to offer substantial services in order to live up to our manifesto for "Live archives" (DAM-LR). While we do see depositors as a class of archive users, depositors have particular needs, for example to prepare and maintain their materials. The kind of interoperability typically provided by federation is based on use of a single SQL-like query to interrogate multiple repositories, which is centred on the information seeker rather than the information manager, which depositors are becoming. MPI's Lamus is a tool that is offering support in this direction. Another concern of depositors, to attain recognition of archive deposits as significant intellectual contribution on par with conventional publishing, can be greatly aided through successful federated dissemination of materials.

Finally, federation allows us to pool and share our strengths, for example, MPI's IMDI infrastructure and programming strengths, INL and Lund's corpuses, and SOAS' expertise in endangered languages.

4. Federating the domain

The goal of federationT is interoperability, the effectiveness of which is traditionally evaluated by the information retrieval measures precision and recall. Precision and recall are improved by using constrained metalanguages. The more lowest-common-denominator the approach to descriptive metadata (and therefore federationT), the less the specialities of participating agencies are reflected. For agencies that wish to serve users more thoroughly, metadata that drives resource discovery needs to be richer and domain-oriented. However, the mere sharing or overlapping of domains does not guarantee a shared semantics or vocabulary. Colomb (1997) shows that inter-database semantics or metadata mapping is a significant problem, even for simple domains. Agents within a federationT are faced with problems of semantic heterogeneity across their databases. Semantic heterogeneity can be a result of differences not solely between data categories, but between participant's understanding of their meanings, interpretations or usages (Sheth and Larson 1990, quoted in Colomb). It can be about differences in formal data models, system or project goals, or as a result of evolution of these over time.

Language archives face quite different data semantics from business and industry. Business data is anchored in well-defined concepts such as quantification, currency, and product codes; these are clearly-understood abstractions, widely agreed to represent key attributes and whose relation to the real world are not subject to interpretation. Libraries also enjoy conventionality of most of their descriptive attributes: well-understood concepts of author, title etc; in addition, these data are typically provided by authoritative publishers, and, as mentioned above, are available to individual libraries from centralised bibliographic sources.

In this sense, the language data world is a quite distinct one, with its descriptive categories, rather than being predetermined and centrally provided, needing to be derived bottom-up from our widely varied data and methodologies. A nomenclature of linguistics exists, but language data does not consist of measurements or key attributes, but speculative and contestable interpretations.[1] Thus, the apparent paradox that linguistics seems to guarantee non-interoperability arises due to the nature of language data (which is already metadata, i.e. we do not have agreed-upon data that will "ground out" the metadata semantics), and due to other factors such as that

[1] For example, a transcription might be changed as the linguist better understands a language's structures. Chomsky's aim was to lay foundations of a linguistic theory that would ground out this problem but it has not been overwhelmingly influential in our areas.

human languages are different from each other in arbitrarily complex ways and that individual linguists seek to emphasise or differentiate aspects of their data or analysis.

Repositories can federate with varying degrees of retention of their "design autonomy" (Colomb), i.e. different levels of change to their information systems to meet the needs of the federation. This is an important issue for DAM-LR. While all the partner agencies hold language data with common but specialised characteristics (e.g. sensitivity; identifying particular persons; emphasis on sound/video in binary formats), they are nevertheless quite specialised. Indeed for most it is a central mission to make a distinct contribution, manifested by creating new infrastructures (e.g. IMDI in the case of DoBeS); others (such as INL) have areal specialisation, or, like ELAR at SOAS, policy specialisation such as collection and implementation of protocol data. In addition, the nature of linguistic data itself is changing and diverging rapidly as the new paradigm of language documentation (a response to language endangerment) grows. For DAM-LR, some concepts are likely to be especially difficult to unify across partners, especially those related to granularity, such as the meanings and cross-mappings of bundle, collection, session etc., and categories of access rights.

5. FederationA: organisational and strategic aspects

The key to dealing with the issues in the preceding section is that the standardisation that enables federationT "is not primarily a computing process" (Colomb); it requires people-based structures, communication channels, and significant resources to maintain these and to enable these to be harnessed towards effective and ongoing development. It is the task of these federationsA to create and host an ongoing, evolving universe of negotiation, knowledge models, and transactions, not merely technical interoperability of terms.

Agencies aiming to form a federation need to be clear about a number of matters, from the semantic ones discussed above, to their purpose and scope, membership, and other strategic, organisational and legal questions. Purpose and scope could range from very broad[2] – to very narrow e.g. 17th century American visual culture (Ninch 2000). These in turn help to create informed and realistic user expectations; i.e. the federationA aims must provide both a forum for sharing and negotiation and a vehicle for disseminating. A co-ordinating body is needed to provide this forum, and to make decisions and strategy, especially in a period of rapidly advancing technology, and where the technology influences what services are expected and provided to users, and have significant financial implications for members.

Therefore, the core of federationA consists of a membership, and its goals. This is totally unlike perceptions of a federationT that consist only of technical standards broadcast from a central agency (the same lesson was learnt in the early development of Z39.50). One could go as far as requiring some form of membership even for users, who must ultimately (for a specialised domain) become part of the community of understanding of the federated metadata and its relationships.[3]

We do have special concerns. For example, conventional authentication systems (such as Shibboleth) exchange minimal data about users, and leave detailed gatekeeping up to individual repositories handling access. However, many linguistic resources have access conditions that associate resources with users, rather as if particular books in a library are not only borrowed under different terms by staff and students, but may be only borrowable by particular named individuals. In our specialist and changing area, federation is not only about searching multiple repositories but about identifying a range of user groups and their needs. This in turn will be enhanced by experience and feedback that a federal forum can incorporate into ongoing strategy.

6. Resourcing and legal aspects

Federation inevitably involves standards, which means formulating rules about implementing them, and, in turn, enforcement through either "incentives or penalties" (Colomb). The mechanism of membership needs to be clear, so that members are signatories to relevant statements of practice, with and formulations of what counts as compliance. Depending in the scope of its activities, a Federation might also be responsible for compliance (and reporting) with various legal requirements (such as data protection, privacy etc.) on behalf of members. These various requirements – heightened by the specific sensitivities and potencies of our holdings – mean that initial statements about trust, ethics etc need to be roundly discussed and formulated as a code to which members assent.

Some of our specialisations create limits to the extent that repositories can be federated. For example, one way of making two data sources comparable is to lose some specificity of the more constrained field – i.e. a "lossy" merge that nevertheless allows users to retrieve the relevant data under most queries. However, where a data attribute has legal or ethical implications (e.g. related to intellectual property, access restrictions, or privacy), then the option to manipulate the appearance, content, or granularity of such data is not open. In

[2] Which can raise problems, such as OLAC's adoption of DC-type scope while appealing to language endangerment for its motivation, thus diffusing its clarity of purpose.

[3] Although we should try to avoid the abuse that the term 'community' currently suffers, such as the "Windows user community" or the "open-source community".

this example, one can see that ultimately federationA is inseparable from federationT, because technologies must reliably implement the policies of members as legal entities with legal and ethical responsibilities and liabilities.

A federation will need a forum or body that can answer the questions that a legal mind will ask; questions such as: Who owns what? What are the risks and who is responsible for them? Where are the boundaries between agencies? How are differences across jurisdictions handled? Who is accountable? Who can communicate on behalf of the federation? For example, privacy legislation require that someone meet an individual's requests to examine data held about them, which would need to be handled initially at the same level as the "seamless interface" that federationT implies to the wider world. Ultimately, such legal and organisational aspects probably need be formally modelled and integrated into the implementation – again, we see the co-dependence of federationA and federationT.

The activities described in sections 4 and 5 above cannot take place without resources. However, in some cases, the resource base can be hidden or go unnoticed; for example, where participants are (a) performing tasks that are part of their core remit i.e. for which local resources can legitimately be expended; (b) public institutions such as libraries that are expected to develop public infrastructure; or (c) in a homogenous, stable, and well-integrated domain, so that benefits from investment could reasonably be assumed to accrue to all participants. Many of these conditions do not hold for the DAM-LR partners and their domains. Therefore, developments are dependent on obtaining sources of funding together with negotiations about the dedication of local members' resources to the federation's benefit. Again, this will place constraints on the processes for membership.

People resources are also needed: Ninch suggest that a federation may need access to a number of types of skills not only on the IT side (e.g. systems analysis, user interface, programmers) but also linguistic, archive, IP and legal experts, representatives of user groups.

7. Conclusion

DAM-LR is providing a useful testbed for the development of a federation of language resource archives, which could be extended to other nascent groups, such as DELAMAN. It already meets several of the considerations discussed above; in particular, we (i) have clear and constrained tasks and membership; (ii) there is a project and funding scenario within which our tasks are negotiated and resourced. On the other hand, it would be misleading to ignore the diverse and distinct organisational, strategic and implementation issues, and to conflate them all under the one term 'federation'. This paper has shown that a federation will weave together aspects of federationA and federationT.

The function of a federation, then, is to:

- supply services to particular communities (cf. OAI "designated communities")
- to supply those services from allocated resources, i.e. federations must *choose* the communities they will serve (for which there needs to be a forum for negotiation and evolution)
- supply services that take advantage of its members' resources, priorities and values
- to manage its membership and resources in support of the above

References

Colomb, R.M. 1997. "Impact of Semantic Heterogeneity on Federating Databases" *The Computer Journal* Vol 40, No. 5, pp. 235 -244.

DAM-LR (partners). 2006 *Live archives.* Pamphlet.

Dempsey, L., Russell, R., Heery, R. 1997. "Discovering Online Resources. In at the Shallow End: Metadata and Cross-domain Resource Discovery." http://ahds.ac.uk/public/metadata/disc_07.html

Lynch, Clifford 1998. "The Dublin Core Descriptive Metadata Program: Strategic Implications for Libraries and Networked Information Access." In ARL 1998 (196), Association of Research Libraries

NINCH 2000. "Federating Digital Image Repositories and Interpretive Information." *http://www.ninch.org/bb/proposals/visual2.html*