Culture Change in Data Management

Peter Wittenburg
Technical Group, Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
Peter.Wittenburg@mpi.nl

1. Introduction

Rather stable workflows that have been established in the research domain over hundreds of years are currently being questioned due to the enormous technological innovation. Still many researchers maintain their private data backyard and are not willing to make their data explicit and therefore sharable. However, for several reasons the pressure is being increased to deposit data in trusted centers and in doing so making it accessible to others, taking care of its long-term persistency, making them citable, allowing verification of resulting publications and enabling community based enrichments.

A simple deposit strategy will not be sufficient, since it could end up in a cemetery of data that cannot be used for many reasons. It would mean that resources could not be found easily and usability could be hampered by unverified content. Thus we need guidelines for proper data management that covers the whole life-cycle of data to ensure usability. The success of the emerging e-Science scenario will largely depend on the compliance with these guidelines.

We will first discuss a number of dimensions that need to be considered when talking about proper data management. Based on these we will derive a number of basic principles that should guide our future workflows when creating, adapting and consuming research data.

2. Dimensions of Data Management

2.1. Data Deluge and Metadata

Most researchers also in the humanities are confronted with an increasing amount of data objects which they either create themselves or obtain from others. A researcher documenting an endangered language together with a few collaborators reported about having more than 5000 files (sound, video, annotations, lexica, field-notes, several versions of each, etc.) on his notebook which he could not manage anymore in the usual ways using filenames and directories. From his own experience he understood that he needed to maintain for example metadata that in detail describes the resource, its provenance and its relations with other resources. But of course the metadata which people are creating is dependent on the type of resource and the research purpose they have in mind. Some researchers are happy to define the resource with the help of a few elements; others who want to do longitudinal research differentiating between the genders of the recorded subjects need to also add age and gender information. And of course annotated media files need to be described by different elements than a lexicon for example. For a repository with currently 50 Terabyte stored in about 1 million of objects maintaining high quality metadata is a must. Creating metadata directly at the beginning of the life-cycle is an urgent requirement.

2.2. State of Data

Data kept on a notebook for example does not have an explicit state, since the researchers can do any kind of manipulation at any time. Only data that is handed over explicitly to a trusted repository that takes care of associating a unique identifier, does time stamping and will check authenticity, has a defined state and is thus citable. As we all know, however, certain data types will never be finished such as for example lexica. There are always new words or new findings about certain words which need to be added. This and the chance of having overseen errors is the reason why still many researchers keep saying that it is too early to make data explicit. On the one hand we need to change culture and convince researchers to submit an imperfect version and on the other hand we need to make sure that the repository has mechanisms to add new versions which then will have their own identity and thus being citable as well. It is a matter of policy when a new version is being uploaded from a private workspace. Experience has shown that if researchers do not have such an active deposit strategy, most of the data will become lost due to a variety of reasons. This does not have to be caused only by a distortion of the local hard disk, but may also be caused by too high costs to convert resources into appropriate formats.

2.3. Data Enrichment, Curation and Costs

Research data is dynamic in so far as they are integral part of research workflows. New research challenges will lead to modifications, extensions, annotations and new relations and it is hardly predictable how research interests will develop. Extensions can be at the collection or resource level, in the first case metadata acting as glue to capture the relations has to be extended and in the latter case new versions need to be created to keep the original version citable. Whatever enrichments are carried out an immediate curation towards proper standards and quality control needs to be recommended to keep the costs of data persistence and accessibility low. Studies have shown that the later the transformations to agreed standards and to quality enhancements will take place, the higher the costs of data management will be. Of course every transformation of content as a result of a curation step will manipulate the content that can even lead to interpretation differences. Therefore, each repository needs to state explicitly what its policy is with respect to the preservation of the original versions and possible effects of transformations.

2.4. Granularity, Identity and Authenticity

In the distributed e-Science scenario where data objects need to be replicated for various reasons it is important to identify each of them by a unique and persistent identifier. The unique identifier and some associated information are representing the object. Essential information associated with a persistent identifier is:

- a checksum which allows to proof authenticity of the resource (is it exactly the object that is expected)
- information that points to the different copies (where can one access the object)
- restricted metadata information that immediately helps the human user to identify the resource (information useful for citations)

It is the task of the repositories to offer precisely the expected resource, once a PID is resolved to an access path, i.e. that every new version needs to be associated with a new PID. One obvious consequence is that each repository needs to have a policy about the granularity of the stored objects. An instance of a database management system covering many types of annotations of a multitude of media recordings and lexica for example is a very complex object and is subject of continuous changes. Linguistic identity then would be embedded in a container, i.e. identity management needs to be done within the database container by the embedded application logic. This can lead to severe data management problems in the long run. It seems to be much more optimal to store linguistically defined units as separate entities at file system level and associate them with PIDs, metadata and provenance information to allow proper data management. The citer proposal for the ISO 24619 standard describes the advantages in more detail.

2.5. Context and Aggregations

In general the interpretation of data objects can only be done by taking contextual information into account. An annotation can only be interpreted comprehensively when the annotated resource is available, a resource often can only be analyzed when metadata containing additional information about the resource can be analyzed as well, etc. Context is manifested by relations to external objects and the relations can have heterogeneous types. Dependent on the type of the relation there are different ways to store and manage them. It is widely accepted that metadata at aggregation level is used to indicate that a number of referenced resources belong to a certain thematically grouped collection. Some are using file headers to indicate the relation to other resources. However, this concept combining information of completely different type is not as generic and often requires redundant information management which is also error prone. The advantage eventually is that information can’t be lost as easily. Aggregations are recursively defined and therefore easily allow building hierarchies which can be used in various ways to simplify access and management.

2.6. Preservation and Interpretation

In general data resources need to be interpretable over a longer period of time for various purposes; those belonging to our heritage for example should be accessible without any time restriction. For management reasons it is widely accepted to differentiate between the pure bit-stream and all layered information such as type of technical encoding, structural aspects and semantic encoding. For preservation purposes it is of primary relevance to make sure that the bit stream information will not be changed and that it can be extracted from the carrier. Given modern storage technology this can only be achieved by a) regular migration to new carrier technology with the help of automatic procedures to make it feasible and b) distributing the data to maintain a number of copies. Long-term interpretability is dependent on the support for all other layers of encoding. Also in these respects we need to consider technological innovation and therefore carry out transformations of the content whenever new standards are emerging. As indicated above, this can lead to changes of the content. A good example is the need to transform video streams to support new video codecs. Concatenation effects can even lead to artifacts influencing interpretability. To understand potential risks it is of great importance to store provenance information so that experts later know what kind of manipulations have been carried out on the bit-stream.

2.7. Replication and Synchronisation

As indicated, preservation of data requires creating replications at different locations. The replication process must be safe which can only be guaranteed by checking authenticity after each operation. Replication assumes a pure master-slave relationship, i.e. data is only copied in one direction. In addition data can be copied by ignoring or including context. Currently, most replication is done by duplicating the bit-streams per resource. Only replication at logical level will include contexts and consider relations. Container formats such as METS for example cater for this kind of packaging, but they often fail since they cannot cope with the dynamics of research data. Replication, however, is not sufficient when not only the master resource, but also the replications are included in researchers' workflows which may result in a bi-directional traffic. A careful synchronization must take place in this case solving more complex versioning aspects.

2.8 Interoperability and Standards

Language resources are created in a distributed fashion by many different creators using tools optimized for efficient generation. The result is a highly fragmented landscape of resources at this moment which cannot easily be combined to coherent virtual collections or being processed by the same analysis tools. In the e-Science scenario, however, we need to overcome all boundaries hampering flexible and seamless combination of resources and tools. The only way to overcome this fragmentation is to promote and rely on widely agreed standards that are certified by recognized standardization organizations or initiatives such as ISO and TEI for example. Relying on standards for technical encoding (UNICODE, linear PCM, MPEG, etc.), for structural description (schema-based XML) and linguistic encoding (tag sets etc.), however, means that active existing communities applying certain best practices would be excluded from participation. Converters need to be available as web services to dynamically offer such resources compliant with the standards. In the case of static resources a one-time conversion may be appropriate.

2.9. Download-First and Ecology

Researchers often don't trust the availability of Internet-based services when it comes to the core of their research work. They want to have all data immediately available in all kinds of circumstances and want to have a secure workspace which is well protected. Therefore, much data is copied again and again due to the download-first paradigm mostly researchers are still relying on. For annotations including a few bytes this does not really form a problem, but when larger corpora or media streams are involved this requires considerable network and storage capacity. Given the increasing amount of data we recognize that this culture will need to be changed to meet the requirements of green computing.

2.10. Sharing and IPR Issues

Resources are increasingly often accessible via the Internet and promote the sharing and combination of resources to virtual collections. The Open Access initiative is working hard to convince researchers to provide their data which has been created by governmental funds to everyone. In most cases one can speak about a win-win situation, since researchers contributing to the open resource domain will take profit from the contributions of others and can theoretically work on much larger sets of data that in general should result in theories with a better empirical grounding. Yet, researchers are often not willing to share resources for various reasons that can range from protecting knowledge, adhering to license conditions up to protecting information about persons who were recorded. Also in this respect we expect a change in culture although personal rights and ethical constraints must be taken serious. Obviously we need a simplification of license conditions for researchers to make access to the many resources practically feasible.

We can indicate an interesting paradox, however: many researchers are not hesitating to give away ALL rights on their data very easily when they make use of the offers from Google, Amazon, YouTube etc, but hesitate to deposit their data in trusted research-oriented data centers.

2.11. Quality Assessment

Trust of researchers in repositories and the accessibility of services can only be achieved stepwise by offering services at a certain level of quality. As it is already the case in high energy physics we need to assess the service levels offered by repositories and other service providers. In fact two assessment procedures are now very well-known: TRAC which is becoming an ISO standard and Data Seal of Approval which is lightweight procedure and therefore more appropriate for our domain. We need to change culture also in this respect to ensure that service providers undergo these quality assessments, since otherwise we will hardly be able to change culture.

3. Basic IT Principles

Based on frequent discussions about the mentioned dimensions in proper data management during the last years at many conferences and workshops a number of basic IT principles have been identified which we need to adhere to when we want to improve the situation towards an e-Science scenario.

3.1. Atomic Objects

The creation of language resources is driven by research questions or potentials. National speech or text corpora for example are created to have a set of resources that are uniform according to a number of design criteria. Increasingly often, however, resources are combined by users due to criteria which are different from those the designers had in mind, i.e. they are re-purposed to answer unforeseen research questions. To allow re-purposing it is important to not only give access to collections, but to each singular resource object. High granularity is a well-known principle in IT to allow unrestricted combinations. In the e-Science scenario no one can predict how resources will be used. It is the task of the communities to specify what exactly the level of granularity is. Different types of users will come to different conclusions. For most linguistic operations a lexicon is the unit that will be addressed, for semantic web researchers it will perhaps be the implicit assertions hidden in a lexicon that can help in reasoning. For annotations it could be all annotations at a certain linguistic level (morphology) that is created by one coherent step of operation. Certainly, it will be one of the major tasks of a discipline such as linguistics to come to recommendations about suitable units. Proper resource models will allow accessing its parts. Certainly also that one resource should not include different information types (video, audio, texts, annotations, lexical attributes, etc) but that they should be maintained externally. This requires, however, stable references for example and a proper container model.

3.2. Explicit Syntax and Semantics

One of the major requirements to achieve long-term interpretability is to exactly specify the structure of the objects used and to register the structure specification so that everyone can easily access them to extract parts for example. Currently still this aspect is not taken serious enough. Many resources are being created with the focus on optimizing creation time, i.e. specific tools are used that do not require adhering to well-specified XML schemas. Inconsistencies can be expected and in general conversion costs much curation time. We need to move towards a culture where for research data certified tools will be used that support the use of registered and well-known schemas such as for example Lexical Mark-Up Framework (LMF). Even more difficult is the transition towards declared semantics, i.e. all tags that are used by linguists to encode linguistic phenomena should be taken from an open concept registry such as ISOcat or at least refer to such concepts. This is the only way to make slow progress towards semantic interoperability. Of course it is known that there is no widely agreed descriptive linguistic system which has to do with a number of reasons such as in particular the large differences between the languages of this world and its non-linearities. This implies that there will be disagreements with respect to the definition of many of the concepts. We also need to accept that communities are already using a variety of tag sets such as STTS and EAGLES for morphosyntactic description for example. They need to be supported which means that mapping systems need to be worked out which are not at all trivial to establish. Also in this respect ISO should come with a widely accepted solution.

3.3. Persistent Identifiers

The e-Science scenario will depend on millions of stable references and their machine resolution. Repositories and registries need to identify resources, services, concepts, schemas etc. Currently we cannot claim that the URLs which people are creating all over the world are stable enough. The W3C TAG is suggesting using cool URIs, but in reality still people are using in most cases URLs, i.e. URI references that are including semantics, physical path names etc. Reality has shown that these cannot be maintained over many years with some exceptions. ISOcat concepts for example are referred to by cool URIs which makes sense since ISO will guarantee its existence. Being faced with this dilemma many institutions and companies are now registering URNs (libraries) and Handles (DOI, EPIC) as an explicit step. However, another level of complexity is introduced which requires maintenance and organizations that can be trusted for their long-term commitment. The advantage of Handles for example is obviously that they can be associated with additional information such as paths to several copies, citation information, checksum etc. without losing resolution efficiency.

3.4. Formats

We need to distinguish between formats that have different functions: (1) Most important are formats that are used to archive resources. They need to be interpretable for a long time and therefore strict adherence to open standards is essential. Encapsulation for example in a database system cannot be recommended. (2) Of similar importance are exchange formats. Since the requirements are similar, they are normally identical with the archiving formats. (3) Work formats are internal and chosen for operational efficiency. They don’t have to be documented and cannot be subject of archiving or open exchange. We need to create a culture, however, where internal formats cannot be seen as primary formats for research data. First, a proper schema-based archiving/transfer format needs to be defined compliant to standards where possible and then for computational purposes internal formats can be derived.

3.5. Standoff

It is almost needless to say that the principles of stand-off should be maintained in case of annotations, since the original information should not be modified.

4. Conclusions

In the emerging e-Science scenario users should be able to easily combine data resources and tools/services; and machines should automatically be able to trace paths and carry out interpretations. Users who want to participate need to move from a down-load first to a cyberinfrastructure paradigm, thus increasing their dependency on the seamless operation of all components in the Internet. Such a scenario is inherently complex and requires compliance to guidelines and standards to keep it working smoothly. Only a change in our culture of dealing with research data and awareness about the way we do data lifecycle management will lead to success. Since we have so many legacy resources that are not compliant with the required guidelines, since we need to admit obvious problems in particular with standardization in the area of semantics and since it will take much time to establish trust at the side of researchers, the e-Science scenario can only be achieved stepwise which will take much time.