

LDGrid: Requirements and Technologies

John Gekas, Udo Kruschwitz, Simon Lucas
Department of Computer Science
University of Essex

Hershbinder Mann, Dan O'Neill
Department of Health and Human Sciences
University of Essex

Department of Computer Science
University of Essex
Technical Report CSM-444
ISSN 1744-8050

December 2005

1 Introduction

Health data for research and policy are distributed amongst numerous organisations and rarely have consistent structures and meaning. The domain of learning disabilities (LD) is a particular case where relevant data sources are distributed amongst government agencies (health and local authority), voluntary sector organisations and academia. The *LDGrid* project is broadly looking at how learning disability-relevant information could be made more accessible and useful at both the strategic and individual levels.

Here we report on a workshop that was organised as part of the *LDGrid* project. The aim was to discuss possible technical solutions for increasing the accessibility and usefulness of existing learning disability-relevant data. We looked at the following aspects in particular:

- Identification and examination of existing data and information sources, both individual level (from surveys and administrative sources) and aggregate
- Building up a firm understanding of the user requirements both for policy and research
- Examining the technical feasibility of joining up disparate data sources within the service providers using *grid* and *Semantic Web* technologies.

At the workshop there were two groups of participants. The first consisted of computer scientists and informatics specialists, and they addressed the issues from a purely technical perspective. The second one comprised representatives of groups that might want access to the information, whether for research, service planning, or to help support individuals who have learning disabilities. Initially, the two groups considered the issues separately; this was followed by a joint session discussing material arising during the technical and the user sessions.

2 User Requirements

Participants in one stream of the workshop were representatives of stakeholder groups who currently access and use information about different aspects of learning disability and learning difficulty.

The interests of those who contributed were in keeping with the concept of a requirement for accessing information at different levels, from aggregate, population level, through to information relevant to individual cases. This included representation from:

- The OECD (Organisation for Economic Co-operation and Development, Paris): *Students with Disabilities, Difficulties and Disadvantages - Statistics and Indicators for Curriculum Access and Equity (Special Educational Needs)* - The DDD project.

The project seeks to improve educational and social outcomes for students with disabilities, difficulties and disadvantages (DDD).

- Essex County Council: Information needs for planning, service management and research.
- Colchester Primary Care Trust: Information management for health.
- The Northeast Essex Inclusive Communication Project. This is a five-year project, supported by the Learning Disability Development Fund that is working with individuals, statutory service providers and the wider community, to develop awareness and skills to meet the communication needs of people with learning disabilities.
- SAFE (Support for Asperger Families in Essex). This is a voluntary organisation that provides information and support for those affected by Asperger Syndrome.
- Clinician / practitioner viewpoints.

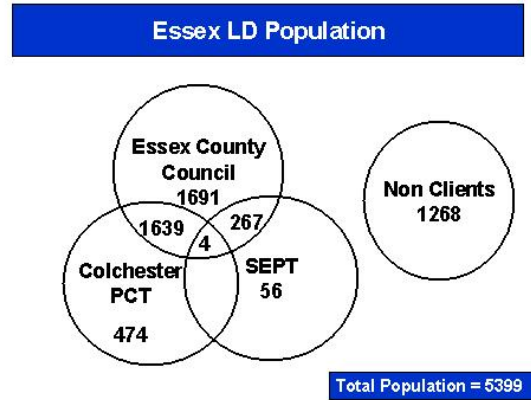


Figure 1: People with learning disabilities known to services in Essex

2.1 The Local Context

Participants provided a valuable insight into the local context, including an illustration of the difficulties inherent in collecting and maintaining reliable data about people with LD who are receiving services from health or social care providers.

Figure 1 is derived from data provided by the TABBS learning disability register, and illustrates the number of adults in the county of Essex who are known to one or more of the statutory sector service providers (PCT = Primary Care TRust, SEPT = South Essex Partnership Trust). The figure refers only to statutory services, and the client or patient base that they support. It is thought that many more people with a learning disability are living in the community and receiving support from without the statutory sector. The publication of the government Green Paper *Independence, Well-Being, and Choice* [4] places the responsibility on local authorities to develop preventative services, so people do not need to access statutory specialist services.

Therefore, one major concern is that of trying to identify people with learning disability who are not currently known to the statutory services so that the demand for preventative services can be quantified.

2.2 Data Landscape

Some understanding of the learning disability information space had been developed through the initial literature search and previous contact with stakeholders (see Figure 2). This identified data silos at different levels, ranging from those held at a local level (for example, the records maintained by General Practitioners), through Regional data holdings, to National data sources, such as large-scale surveys and census data. At each level, these data can be generated and held by actors from different sectors - statutory, voluntary and private. It is generally thought that there is little or no sharing of data across sectors.¹ Within each sector there are then a number of agencies and organisations with variable degrees of data compatibility and exchange. However, the discussion in this workshop generated a far more detailed picture of the complexity and the issues raised for the field of learning disabilities. In particular, contributors from both health and social services highlighted the fragmented and partial nature of data held within organisations as a source of continuing difficulty.

Figure 3 illustrates that, while there is a degree of data sharing between Essex County Council, Colchester PCT and South Essex Partnership Trust, the County Council also uses (or would like to use) data from a wider range of sources. A notable example is the Employment Service, where it is considered that data sharing, for example to ensure that individuals are receiving the correct levels of benefit, could be extremely beneficial.

¹One exception to this would be where statutory bodies commission services from providers in the private sector, when they are able to insist on certain data being made available for audit and management purposes.

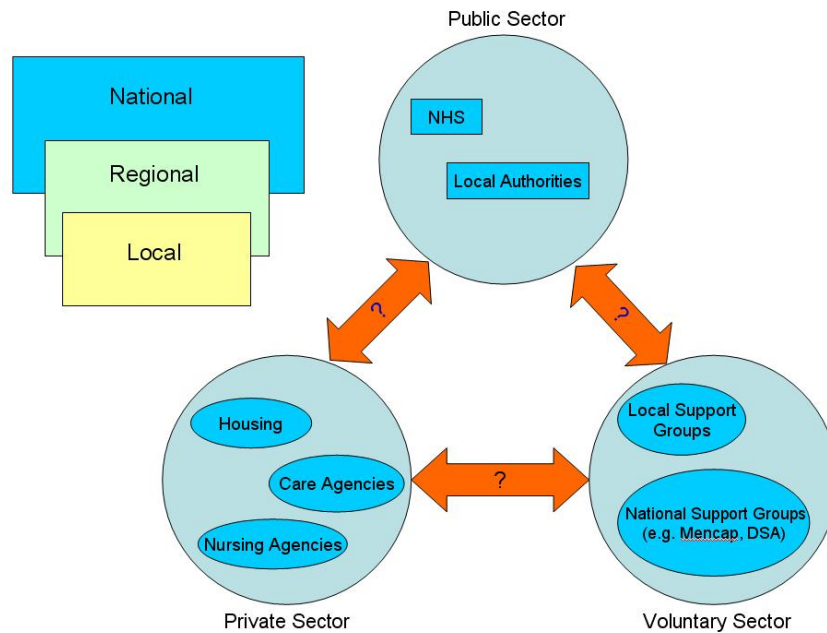


Figure 2: Data landscape

Information needed and currently collected by Essex County Council includes the following:

- Basic demographic details: *Age, Sex, Ethnicity etc.*
- Type and level of disability: *Down Syndrome, Autistic Spectrum Disorder.*
- Identified support needs
- What services people are using: *how often, and how much they cost.*
- Where people are living
- What type of accommodation: *Family home, Tenancy, Registered Care.*
- Performance monitoring information: *how many people have jobs, use Direct Payments, registered with a GP.*

2.3 What the Users Want

In discussing the existing information space and the associated difficulties, participants began on the task of identifying answers to the practical problems they faced. On the basis of the criticism levelled at the data landscape as it presently confronts them, the users in the workshop considered the characteristics of a technical solution that could counter those same difficulties. In producing a general model of 'joined-up' information sources, the overriding themes may be summarised as the following:

- *Integration:* There is a clear demand for a solution that can tackle the problem of data fragmentation and dispersal. The fact that data are held by multiple organisations in multiple formats is a critical issue that could be addressed by a system that allows access to key sources from a single point of entry.
- *Efficiency:* Data relating to individual persons/cases may be held on multiple databases within the same agencies as well as between them. Users pointed up the fact that if communication between these databases is not permitted, time is lost to data entry and the use of that data may itself become inefficient as data replication arises. Moreover, information can quickly become outdated if the various sources are not updated concurrently as changes occur. For example, a change to an individual's address might be applied in one database but if this is not known to another agency, any associated sources will also become inaccurate.

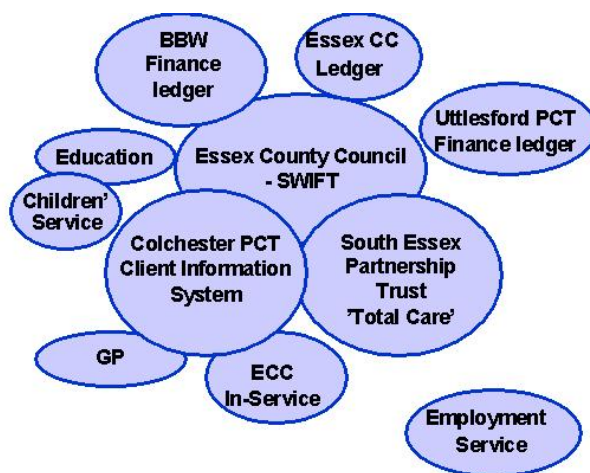


Figure 3: Essex social services - main data providers

- *Customised Searching and Access*: Information must allow for the planning of services and inform the targeting of resources in accordance with need. Data sources must allow users, such as health practitioners, the ability to access information in such a way that it enables analyses of diverse cases within different time frames, places, and contexts. The level of specificity of any particular query would need to be established by the user in relation to particular criteria.
- *Security*: There is necessary for all data sharing to observe whatever data protection and confidentiality protocols apply. Locally this means conforming to the Essex Trust Charter² as well as greater data protection concerns. Consequently, any proposed system of data sharing must incorporate access control.

3 Technical Issues

“The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.” [1]

The idea of a *Semantic Web* has been around for a few years now and the aim is to move from a purely syntactic description of data (e.g. plain text that includes some information about how to display the text) to something that is much richer, where the documents contain semantic information. Semantic information is already being used nowadays, for example the *meta tags* of Web pages often contain keywords assigned by the author that characterise the page content, information about the author, the creation date etc.

This idea of semantically encoding information is very promising also for projects like *LDGrid*. We simply have to imagine that the “documents” we are dealing with are for example patient records. If such records can easily be exchanged and interpreted by computer programs, we will be in a position to easily join up different data sources and employ them in integrated applications.

However, despite the fact that there is a constantly growing quantity of data sources containing explicitly encoded semantics, most data still comes with much less structure; and the LD domain is no different to other areas in that we will aim at incorporating existing databases and legacy data instead of trying to start from scratch. This needs to be taken into consideration when designing new applications.

²<http://www.essexinformationsharing.gov.uk/index.htm>

3.1 Context

The technical experts at the workshop brought in a lot of expertise in developing and using data representation standards but they also were involved in large scale projects that employ *Semantic Web* technologies, e.g.:

- MIAKT³: *Medical Imaging and Advanced Knowledge Technologies*, funded by the Engineering and Physical Sciences Research Council (EPSRC)
- CancerGrid⁴, funded by the Medical Research Council (MRC)
- DIP⁵: *Data, Information, and Process Integration with Semantic Web Services*, funded by the European Union
- CLEF⁶: *Clinical e-Science Framework*, funded by the MRC.

3.2 Data Representation

One of the key issues in designing a system that makes LD-related data more accessible is the question of data representation.

Despite the fact that we are still far away from Berners-Lee's vision of a *Semantic Web*, there has been considerable progress. Part of the work towards developing the *Semantic Web* has been the development of standards to express and represent knowledge in a way that makes it possible for computer programs to "understand" such knowledge. Some of the most prominent data representation standards are:

- Extensible Markup Language (XML)⁷
- Resource Description Framework (RDF)⁸
- Web Ontology Language (OWL)⁹.

XML is a markup language derived from SGML. Unlike HTML, which comes with a fixed set of tags and encodes mainly the visual representation of a document, XML allows an author to define his or her own tags. So it is in fact a metalanguage. If different authors agree on the same XML tags, then it is easy to exchange marked up documents and interpret these tags in a specific way. XML Schema are a way to describe the structure of an XML document and constrain the elements used in it. However, just XML is not enough to represent *ontological knowledge* appropriately (we will come back to this point further down).

RDF is a very simple data model based on triples. These triples represent a subject, a predicate and an object. Statements in RDF describe properties of resources (objects). RDF is however not rich enough to describe resources in enough detail. One extension to RDF is RDS(S) (which stands for RDF Schema). OWL is an extension that goes beyond XML, RDF and RDF(S) by providing additional vocabulary as well as a formal semantics which can be seen as a descendant of traditional knowledge representation formalisms such as KL-ONE [2].

For many cases, the use of RDF or OWL is unnecessary, and it may be better to mark up data in using domain-dependent XML, with the document types being defined by XML Schemas directly. This is simpler and more compact than using large sets of RDF triples, and has the advantage of better interoperability with databases and other tools. For example, we heard on the CancerGrid project how the use of plain XML allowed semi-automatic web-form creation to be done using the Microsoft InfoPath tool.

Beyond the actual data representation issues the more general question is what information to represent and how to acquire any knowledge that goes beyond the data already available. Significant work has been conducted in recent years in the *knowledge representation* community. Much of that work is closely linked to the idea of building a *Semantic Web* by means of *ontological* knowledge, i.e. highly structured knowledge sources that encode information about the world or about a particular domain of interest. The freely available U.S. *National*

³<http://www.aktors.org/miakt/>

⁴<http://www.cancergrid.org/>

⁵<http://dip.semanticweb.org/>

⁶<http://www.clinical-escience.org/>

⁷<http://www.w3.org/XML/>

⁸<http://www.w3.org/RDF/>

⁹<http://www.w3.org/TR/owl-features/>

Cancer Institute (NCI) ontology (now implemented in OWL) is an example of a large structured knowledge source for the cancer domain.¹⁰

Referring to learning disabilities in particular, no official effort seems to have been made. Some official classification efforts of LD-related concepts and disorders exist as part of bigger classification frameworks, such as the DSM-IV¹¹ and ICD10¹², but still consistency seems to be a major problem: indeed, in the domain of learning disabilities, relationships between causes, disorders and treatments are more vague than other domains of medical informatics, due to the large impact psychological and social factors have.

Ontologies and customised versions of existing language resources like *WordNet* [5] are being successfully employed - for example to search product catalogues and other document collections held in relational databases - a sentiment that was confirmed by participants of the workshop. However, there are some fundamental problems with ontologies. First of all, an initial effort is needed to construct some knowledge in the first place, which is a non-trivial task - despite the fact that ontology editors such as Protégé¹³ have been developed and are freely available. Furthermore, a major problem is that “there is a lack of methods and tools supporting and facilitating ontology reuse” [8]. Current research activity focuses on addressing such problems, for example there are numerous projects at the University of Manchester and at the Open University. Nevertheless, these problems are not likely to disappear in the foreseeable future.

3.3 Security Issues

As pointed out previously, it is difficult enough to share data in the first place and to make existing data sources available to others [6]. Apart from that we will have to seriously consider confidentiality and security issues.

However, security has widely been neglected in the *Semantic Web* framework, e.g. the semantic relations encoded in RDF for example are either accessible or they are not.

If we look beyond that, we find that distributed applications that use *grid technologies* pose an additional challenge in terms of security and data sharing. Such technologies are often considered to involve high performance clusters for fast processing of massive amounts of data. However, in projects such as *LDGrid* the crucial problem is not so much the *amount* of data but much more importantly the *distribution* of these data. This leads to an alternative view on *grid technologies* which focuses on supporting *virtual organisations* by sharing a pool of distributed resources. Often these data collections contain information about individuals and the main concern becomes how such data can be exchanged, shared, joined up applying strict confidentiality and data protection guidelines.

If we adopt such an interpretation of *grid technologies* for projects such as *LDGrid*, then we will find that the central question is how to agree on access policies and security issues when it comes to sharing data.

Often it will be easy to share data but much more difficult *not* to share data, i.e. making different parts of the data accessible to different users by distinguishing a number of user groups, or users with different roles. This is an area that has attracted a lot of attention recently and is very much work in progress. *Shibboleth*¹⁴, for example, is an initiative to address these very questions by developing an open, standards-based solution to allow organisations to exchange information in a secure, and privacy-preserving manner.

3.4 Application Architectures

The idea of the *grid* as a means of supporting *virtual organisations* by sharing a pool of distributed resources is very appealing. Such distributed architectures are becoming more and more common. Often they incorporate dedicated *Web services* serving individual parts of the overall system. This type of architecture also makes it easier to replace individual components and plug in new modules as they become available.

Both CancerGrid and MIAKT for example use such an overall architectural approach. For the data integration step we suggest to adopt such existing platforms and frameworks rather than developing yet another architecture.

¹⁰<http://www.cancer.gov/cancertopics/terminologyresources>

¹¹<http://www.psychiatryonline.com/>

¹²<http://www.who.int/classifications/icd/en/>

¹³<http://protege.stanford.edu/>

¹⁴<http://shibboleth.internet2.edu/>

4 Next Steps

There are a number of different applications that we might want to consider to make LD-relevant data more accessible. There are two elemental use cases that we consider [9]. The first is the policy scenario in which the user is concerned with examining the overall picture. The second is the service scenario in which an individual care programme is being managed.

4.1 Policy Scenario

On one hand are those users who want to access the data as a whole in order to plan services, understand trends or research particular topics. These users are likely to be taking an overview and not be interested in individual cases, even if they wish to have access to the individual level data so as to have complete flexibility to test different hypotheses by carrying out detailed statistical analysis. For them the main concern is to have as much data access as possible so that the population can be served most effectively by well-researched, well-managed and well-monitored policy interventions. In other words the better the information, the better the service can be.

4.2 Service Scenario

The other elemental use case is centred on the individual record. In this scenario the care provider is the user and, ideally, they will want access to data in such a way that efficient individual care strategies can be pursued. For example a patient might need residential care and the care worker will want to

- identify available care provision
- liaise with other agencies with an interest in the patient
- liaise with parents or other guardians
- possibly link to support groups or other informal care networks.

For this to happen, an integrated care record, probably based on the NHS number or other identifier, is required. This is not unproblematic. In practice there are many organisational as well as technical and legal barriers.

4.3 Meeting User Requirements

A central question is what sort of techniques are most suitable for the problems at hand - be it a policy or a service scenario use case. It is very difficult to select a particular approach without a clearly specified data landscape that includes a detailed description of data records. Getting hold of real data is however a commonly reported problem and a major obstacle in particular in projects that deal with personal records.

However, if we leave this issue aside, there are two system designs that should address the user requirements we collected. We will give a fairly high-level description of each of these two designs.

The first system addresses the policy and service scenarios just discussed.

The second system design is a result of this workshop as well as our previous workshop on data sharing and confidentiality [6]. One of the findings of that workshop was that a framework is lacking that has the carer or the patient as a potential user of the services in mind. This is becoming increasingly important given the introduction in recent years of legislation designed to promote the inclusion of disabled people and other vulnerable groups, and to support the rights of the individual in society (e.g., the Disability Discrimination Act¹⁵, Data Protection Act¹⁶ & Freedom of Information Act¹⁷). All the confidentiality and data sharing issues apply equally in the context of learning disability services, but the situation is complicated further by the service users' reduced capacity to engage with the development of data sharing systems.

This sentiment is strongly supported by evidence reported elsewhere, e.g. in an Essex County Council commissioned study [7] for which disabled people were interviewed and which found for example:

¹⁵http://www.opsi.gov.uk/acts/acts1995/Ukpga_19950050.en.1.htm

¹⁶<http://www.opsi.gov.uk/acts/acts1998/19980029.htm>

¹⁷<http://www.opsi.gov.uk/acts/acts2000/20000036.htm>

“Lack of accessible information about Social Services, and the potential support and advice that may be on offer for disabled people, emerged spontaneously as a key issue of concern for many of the people we interviewed.”

There is further support for such an approach when we look at the exemplar domain of learning disabilities in particular. It has been reported that an estimated 60% of adults with learning disabilities live with their families [3]. These carers *“face many problems and challenges. They need more and better information”*. One of the specific challenges identified by the government paper [3] is to ensure that carers *“obtain relevant information about services”*.

The second system we would like to see implemented addresses these very issues.

From a technical perspective we suggest to

- adopt existing platforms that have been shown to work well in other projects such as CLEF, MIAKT and DIP
- develop reusable data sources
- use open standards
- have a particular focus on security and confidentiality.

In terms of data representation and exchange issues we will want to move towards *Semantic Web* technologies. For the immediate needs of the project however, we suggest to encode the data directly in domain-appropriate XML rather than RDF or OWL. Nevertheless, whatever data representation paradigm we are choosing, we need to be aware of problems that the artificial intelligence (AI) community had to deal with for decades, e.g. the expressiveness of knowledge representation formalisms *vs.* their usability and maintainability.

4.4 System Design 1 - Joining Up Learning Disability Data

Figure 4 gives an overview of the first system we envisage. This system design is quite generic in that it represents a combination of policy and service scenario. Individual records are linked, access is then performed on a record level or on aggregated data. The user in this case is the health or social service expert (for the service scenario) or a council worker, researcher or consultant (policy scenario).

The use case scenario for such a system looks as follows:

1. The user poses a request regarding quantitative LD-related information (which can be found in one or more different data sources), in the following form:
 - Information listing
 - Consistency check
 - Data comparison between the different sources
2. The system queries the data sources and attempts to retrieve the requested pieces of information by attempting to match the contents of the data sources with the structured schema of LD-related information.
3. Once the requested information is retrieved, it is presented to the user in the appropriate form (listing, comparative etc).

4.5 System Design 2 - Making LD Data More Accessible

The second system (figure 5) does not follow our distinction of service and policy scenarios because we do not look at individual data records but at other information sources that should be made more accessible to users such as carers and patients.

The scenario looks as follows:

1. The system makes use of already existing online LD resources, regarding:

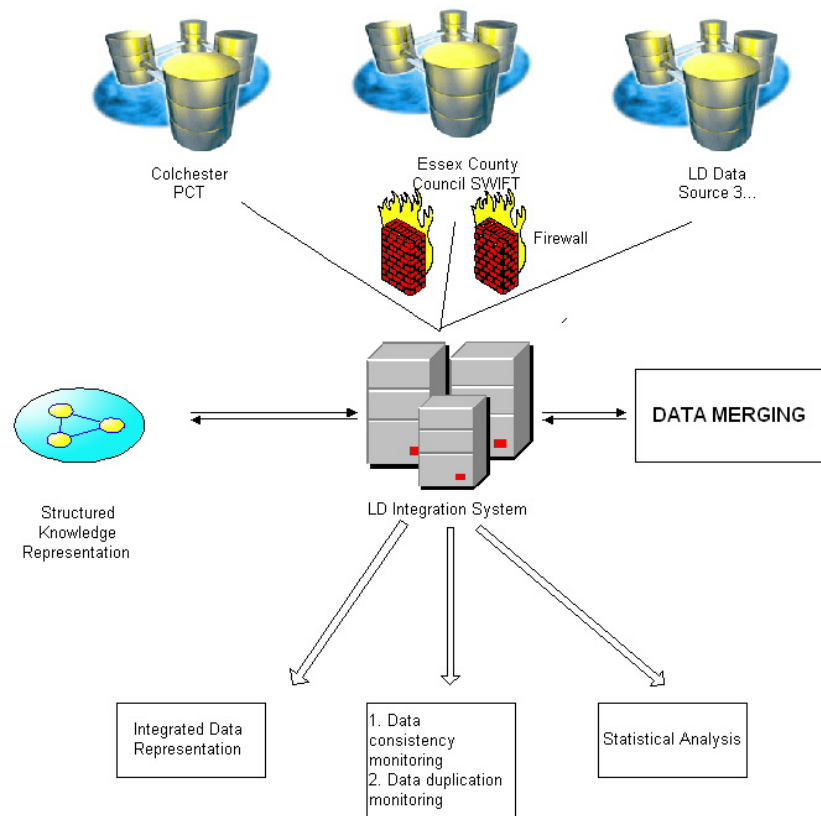


Figure 4: Use Case 1: Joining Up Learning Disability Data

- News and articles within the LD community
 - Papers and articles
 - Qualitative information about benefits, eligibility and application procedures
 - Published statistics
 - Public information on local services
2. Constant monitoring of relevant resources takes place (through the part of the system that is labelled as web crawler).
 3. The contents and updates are semantically annotated based on the common LD-related vocabularies used by the system.
 4. Once the interesting and most relevant parts are retrieved from the text-based sources, these are available for presentation to the users.

Again, this is a very generic system. There are of course a number of issues which should drive the development of such a system. A number of research projects have delivered useful input that should drive such applications in the LD domain, e.g. before the text documents are presented to the user they could be preprocessed to make them more easily accessible by “enriching” them similar to what happens in the MAGPIE¹⁸ project. Such enrichments could be the introduction of hyperlinks that link words to pages that explain these terms further in an attempt to help the readers grasp the content of the documents more easily. These links could also point to manually constructed knowledge sources.

An extension to this system design is an additional dialogue component that guides the user through the space of available information. This part, however, is a whole research area on its own. Current dialogue systems are heavily restricted in terms of domain coverage and they usually access well structured databases.

¹⁸<http://kmi.open.ac.uk/projects/magpie/main.html>

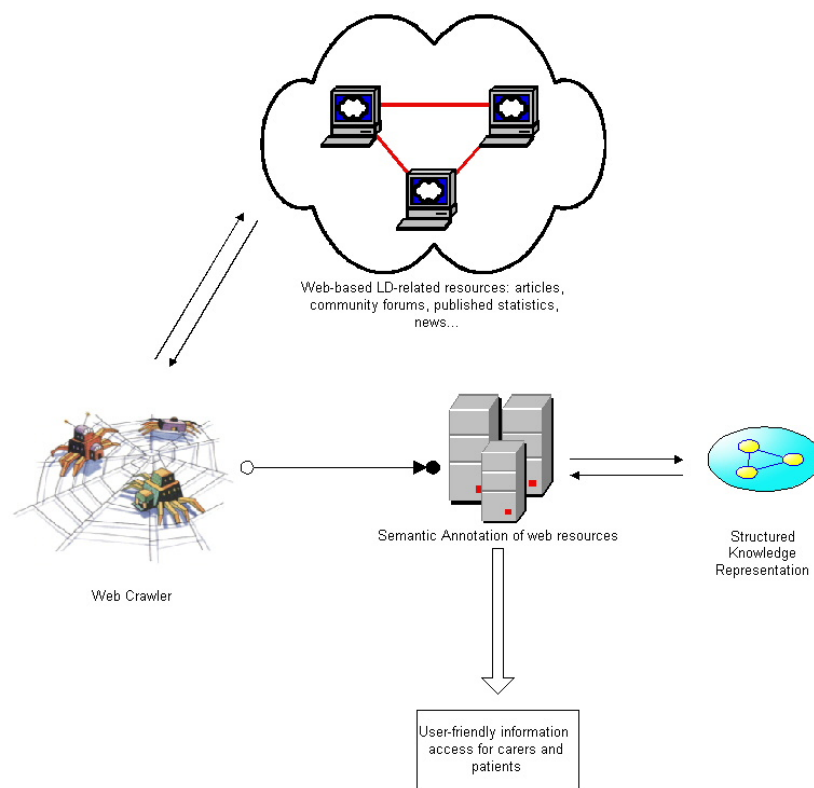


Figure 5: Use Case 2: Making LD Data More Accessible

5 Conclusions

Concerning the wishes from users and providers of data as to what is really needed we saw a number of findings of our previous workshop on data sharing and confidentiality confirmed.

Two of the most prominent wishes that users had are:

- Accessibility of information for users and carers (with a particular focus on *easy* access for non-experts)
- Joining up data sources that come from different providers to overcome the fragmentation of services (e.g. health and social care data).

We outlined two general system designs that we would like to see implemented. Both these systems address different issues in the LD data management cycle.

References

- [1] BERNERS-LEE, T., HENDLER, J., AND LASSILA, O. The Semantic Web. *Scientific American* 5 (May 2001), 34–43.
- [2] BRACHMAN, R. J., AND SCHMOLZE, J. G. An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science* 9, 2 (1985), 171–216.
- [3] DEPARTMENT OF HEALTH. Valuing People: A New Strategy for Learning Disability for the 21st Century. Government White Paper, 2001.
- [4] DEPARTMENT OF HEALTH. Independence, Well-being and Choice: Our Vision for the Future of Social Care for Adults in England. Government Green Paper, 2005.
- [5] FELLBAUM, C., Ed. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

- [6] GEKAS, J., KRUSCHWITZ, U., MANN, H., O'NEILL, D., AND MUSGRAVE, S. Report on the NCeSS Agenda Setting Workshop on Confidentiality and Data Sharing. Technical report CSM-434, Department of Computer Science, University of Essex, 2005.
- [7] JOHNS, T., NICHOLAS, N., JOHNSTON, J., COOPER, G., CARTER, Z., AND MILLER, P. The equal lives evaluation report summary. Essex County Council Publication, 2004.
- [8] MAEDCHE, A., MOTIK, B., STOJANOVIC, L., STUDER, R., AND VOLZ, R. An Infrastructure for Searching, Reusing and Evolving Distributed Ontologies. In *Proceedings of the Twelfth International World Wide Web Conference (WWW2003)* (Budapest, 2003), pp. 439–448.
- [9] MUSGRAVE, S., KRUSCHWITZ, U., AND O'NEILL, D. Confidentiality Issues from the User Perspective - Lessons from Learning Disability Services. In *Proceedings of the First International Conference on e-Social Science* (Manchester, 2005).