INKA

Internet-Based Knowledge Acquisition and Publication Project
Department of Publication Infrastructure
Scientific Information and Learning
KTH - Royal Institute of Technology, Stockholm, Sweden

Erik Sandewall

Extending the Concept of Publication: Factbases and Knowledgebases

This memo is identified as PM-inka-001, version 2, and will be permanently available on the Publication Infrastructure Experimental Website (PIEX) with the URL `http://piex.publ.kth.se/reports/inka/001/`
Manuscript completed on: 2007-12-19
Manuscript webposted on: 2007-12-19

Related information can be obtained from the following www sites:

The PIEX website:	`http://piex.publ.kth.se/`
The author:	`http://www.ida.liu.se/~erisa/`

Introduction: The Concept of Publication is Extended Due to Information Technology

The basic meaning of the word "publish" is to "make public", but other connotations are also associated with that term. Well-defined authorship is usually assumed, which provides an incentive and a credit for the author as well as assigning responsibility for what is published. Wide availability of the publication and its preservation for the future are also assumed, and in the context of science, it is considered important that a publication has undergone peer review not only for the purpose of quality control but also for providing feedback to the author.

One further feature of a traditional publication is (or was) that it is produced by printing text and images on paper, for example as a book or a journal article. This is no longer necessary, because of the developments in information technology, and it is now generally accepted that the term "publication" can be applied to information objects that are only represented electronically, provided that other criteria are met.

However, this raises the question of whether the use of the term "publication" should still be restricted to information objects that consist of text and pictures, or whether (and to what extent) it is appropriate to extend it to information objects with other kinds of content (such as video or audio)?

Most people accept that audio and video recordings can be considered as publications, if only because libraries already accept such recordings on physical media such as magnetic tapes. It seems natural to consider the contents of that medium as a publication, irrespective of the physical object on or in which it is stored.

Computer software, and in particular open-source software represents a more far-reaching extension of the same concept. Such software, once it has been developed, is made available on publicly available websites such as Sourceforge ([1]), so it is easy to argue that it meets the basic criteria for publications. The open-source software community has also implemented and organized an efficient system for interactive discussion about such software, for example using project-specific wiki systems. These serve the same purpose as traditional peer review, namely, for feedback to the authors, improvement of the "publication", and quality assessment for the benefit of prospective users.

In this article I propose that one other type of information object can also be considered as publications: namely, modules for factbases and knowledgebases. A factbase is a collection of elementary information, for example, the population, geographical size, and name of the capital of each of a number of countries. A knowledgebase is a collection of information with more complex structure, for example, instructions of use for a variety of machines, and advise for diagnosing and fixing faults. There is no sharp boundary between these, which is why we consider them together.

Modules for factbases and knowledgebases are similar to software modules in the sense that they are designed so that users can include them in their software systems. However, such modules differ from software modules in

[1] `http://sourceforge.net/`

that they contain information about something in the world, whereas software is, of course, essentially a collection of instructions for a computer. I shall discuss the advantages of extending the traditional publication concept to this new kind of information objects, as well as the ramifications for publishers and libraries.

A Concrete Example: The Common Knowledge Library

The Common Knowledge Library (CKL)([2]) is a prototype publication website for fact and knowledge modules. The first contributions to the CKL contained simple descriptions of well-known objects. For example, there is one module for 'countries', containing a description of each country in the world; another module for 'languages' containing a description with a few characteristics of each major language, and similarly for universities, journals, publishers, and so forth. There are cross-connections within and (more often) between the modules. For example, each country has one or more official languages; each journal publishes articles in one or (sometimes) more languages, and so forth.

Every such module is represented as a text file with a particular structure, using a markup language similar to XML or OWL. Cross-links are represented by each module introducing a name for each one of the entities it contains, and other modules using the same name for referring to that entity.

The CKL uses the word "library" in the sense of "a collection of documents that are made available to the general public". Since it is a website it is not associated with a particular building, and there is no notion of borrowing and returning specific copies of these documents. Words like "repository" or "archive" could also have been used.

There are two major reasons for adding a module to the library. One is the value of the information that it contains; another, and even more basic, reason is that each module introduces a standardized vocabulary for the objects that it represents. For example, the module for countries determines what name is to be used for each country, in cases where this is not obvious. For example, `Ivory_Coast` or `Côte_d'Ivoire`; `United_States_of_America`, `U.S.A.`, or `USA`?

The contents of the basic modules are simple; their main purpose is to introduce standardized vocabularies. Some of the later modules are much larger; for example, the CKL contains a register of more than 15.000 journals and several thousand publishers, with the relevant links between journal and publisher. Other modules contain more complex information; for example, a formalized representation of various publishers' rules for parallell publication by their authors. This information is given as a few English sentences in free text in the Romeo/Sherpa website ([3]), but in the CKL the same information has been converted to a structured form that is suitable for interpretation by computer programs.

[2] `http://piex.publ.kth.se/ckl/`

[3] `http://www.sherpa.ac.uk/`

Much of the early work on the CKL addresses facts from the domain of scientific publication, but there are also some modules for other areas, for example, one for European airlines. It is our intention to gradually extend the CKL's coverage in several directions, and also to proceed from simple information that has the character of "facts" to more complex information that is better characterized as "knowledge", in the sense that was discussed before.

The purpose of the CKL is to make entire modules available, rather than individual facts. Someone who is looking for an answer for a specific question, such as what is the capital of a particular country, or who is the publisher of a particular journal, will be much better served by existing services, such as Google, Wikipedia, or an open website or another website with a query facility. On the other hand, the CKL may be valuable for someone who is setting up a software system to run a service and who needs the information as a basis for that service.

Suppose, to take a very simple example, that you are building a system for which you need a list of all the countries in the world. You do not want to type it in yourself; there are almost 200 separate countries on the planet. You may think that you can easily find such a list on the Internet. However, even for such a seemingly elementary exercise, it turns out that the lists you can find there have their problems (which are discussed below). The purpose of the CKL is to help in such situations.

One important feature of the CKL is that its contents are freely available for use, just like free software. Each module in the CKL is made available under the Free Software Foundation documentation license ([4]), which means that it can be used freely, but if it, or derived works based on it, is republished then the new publication must also be freely available.

Another important feature is that metadata for the modules are also published. These metadata include authorship, the sources that were used for constructing the module, information about the IPR status of those sources, and more. There is also a system for timestamps and versions: modules may be updated, but each instance of a module contains information about when it was published, and older modules are retained for reference.

Quality control for modules is important; users need an assurance about quality. We consider that there are two major aspects of quality control: formal coherence, and factual correctness. Factual correctness means that the information in the module is a correct representation of conditions in the world; formal coherence means that the module is structurally correct.

For example, let us suppose that a module of scientific publishers specifies the country in which the publisher's head office is located, and this is given as `United_Kingdom` for some publishers and `Great_Britain` for others. The module for countries specifies that the identifier `United_Kingdom` is to be used, and it does not define the entity `Great_Britain` at all. In this case the formal coherence is not satisfied, since the structure specification for the module is that the value of the country attribute for a publisher shall be an entity whose type is country, and this requirement fails for some of the publisher entities. On the other hand, if the module specifies the location attribute as `Germany` when in fact the publisher is located in the U.K., then

[4] `http://www.gnu.org/licenses/fdl.txt`

it fails of course with respect to factual correctness, but not necessarily on formal coherence.

Since one use of the CKL knowledgebase is to assist in the interpretation of additional sources, it must in fact contain a collection of alternative names for an entity. In this way it is possible to substitute different names that occur in different sources, with a unique name that is used within the knowledgebase.

Formal coherence can be checked by automatic means; however, other methods are required to ensure full factual correctness. The notation that is used for complex information within the CKL allows one to specify the structural rules for formal coherence, and the support software contains a service that checks coherence for given modules. The coherence-check protocol is also published on the CKL website; one important aspect of the protocol is that it specifies which version (identified by its timestamp) is used for each module, since the choice of version may be essential for the validity of the entire combination.

Once structural coherence has been achieved, it is also important to ensure factual correctness. What we do at present is, first of all, to declare what sources have been used, so that the user can use whatever information may be available concerning the trustworthyness of these sources. Secondly, we try to document what procedures have been used for interpreting and correcting the source information. We have not yet set up a mechanism for reviewing by third parties; this will be a task for the future.

Comparison with other approaches

There are many websites on the Internet that provide query and browsing services for various kinds of information. If you are looking for the name of the currency in a particular country, or the telephone area code for a particular city, you can easily find a website that will provide the answer. However, if you are looking for a comprehensive file with this information for all the entities in a particular group, then it may not be so simple to obtain it from Internet sources, primarily because the owners of the relevant websites may not want you to have the entire file.

The reasons for this may vary. Many of those who operate such websites earn money from them, for example, from advertising on the site, and it is therefore important that they continue to receive visitors. If someone else copies the contents and makes them available elsewhere, then this may result in a loss of visitors and thus a loss of income. Sometimes, too, there may be a combination of free services and paid-for services on the site, and revenue for the paid-for services may be lost if the underlying database is made available elsewhere.

Owners and operators of such sites may therefore set up a variety of barriers in order to prevent their database being used by others. For instance, the site may be designed so that it is very inconvenient to extract its contents, and it may contain supervisory routines that identify attempts to do this. There are legal restrictions too: the maker or owner of the database may own intellectual property rights for the database itself, quite apart from its contents.

However, even in those cases where the website operator does not restrict access to and use of the information in her or his site, there are still many practical problems. Although the HTML code that is nowadays used for markup on most websites can be very easy to analyze, often it is in fact rather messy. The names of entities can vary to a surprising extent, and it is almost never safe to assume that the same thing has the same name in all circumstances. Misspellings and outright errors can occur even on websites that one would expect to be authoritative.

It seems to me that this is why the vision of the "Semantic Web" has failed to materialize. The vision was that information that one author had put on the world-wide web should be usable by other *software* to execute operations, and not merely by other individuals for the purpose of browsing. The Wikipedia describes it as follows:

The Semantic Web is an evolving extension of the World Wide Web in which web content can be expressed not only in natural language, but also in a format that can be read and used by software agents, thus permitting them to find, share and integrate information more easily. ([5])

The first proposed means of moving towards that end, was that web pages should be written in XML ([6])(a new language that is intended, among other things, for defining web pages) and not in HTML (the older language), for two main reasons: it is easier to analyze than HTML, and XML allows the author of the webpage to insert additional information that can be used by visiting computational processes when they harvest the information. Additional languages and other enabling technologies have been proposed later.

The problem is that these enabling technologies only solve a part of the problem, and it fails to address two very important issues: providing motivation for the authors of web pages, and providing quality assurance. These problems can be solved, however, through publishing factbase and knowledgebase modules, as demonstrated in the Common Knowledge Library.

If a webpage is to be useful as a source for computations elsewhere, its designer must do a certain amount of work to make it usable. There therefore has to be some benefit to the author if he or she is going to put in the extra work. If knowledge modules can be considered as publications, then that benefit can come from recognized authorship and from citation, in just the same way as is customary in the academic world.

With respect to quality assurance, the CKL shows how it is possible to set up a system where knowledge modules are provided with metadata and with structural checks. In order to be effective, these checks must operate on a collection of modules, and not merely on a single module, since cross-connections between modules are important. This is feasible in a library-like situation, where one can control and administer successive versions of each module; however, it is very much more difficult if the module information is located directly on world-wide web pages or in dynamic resources that can be updated independently and without coordination.

In addition to the XML-based approach to the semantic web, there has also been much work on other notations that go further in the direction

[5] `http://en.wikipedia.org/wiki/Semantic_web`

[6] `http://www.w3.org/XML/`

of semantics, in particular the OWL notation ([7]). Unfortunately, although this work has resulted in a lot of software development and in numerous ontologies, there seem to be few results that provide useful collections of facts. (An ontology is a classification structure that specifies what concepts are specializations or generalizations of what other concepts). For example, the module library of the `Protégé` system ([8]), which uses OWL as one of its two markup systems, contains a number of different modules for ontologies, but very few modules with information that conforms to those ontologies.

The dbpedia website ([9]) contains a large database that is derived in most parts from the English and German language versions of the wikipedia ([10]). This information is far superior to most other sources by being more comprehensive and by being more accurate with respect to the form and the spelling of names. However, it is based on information that was prepared for being seen by a human reader, and not primarily for automatic processing, and a considerable amount of interpretation is therefore necessary for this source as well.

Value-Added Provided by CKL Modules

The total number of entities in the modules stored in the CKL is approaching 50,000, and is increasing rapidly. Very little of this information has been typed in directly by us; almost all of it has been obtained from open sources that are already on the web, or from sources whose authors are glad to share them. This means that, rather than using a module in our library, it is perfectly possible to go directly to the original sources. However, this would involve repeating the preparatory work that we have carried out before publishing a module. We refer to this work as *interpretation* of the sources, and it is in fact the added value that the CKL provides to its users.

In many cases, the original form of the source information is a table in which each line represents one entity in the class under consideration, and each column contains an attribute of that entity. The table may arrive as an Excel sheet, a table within an HTML page, or a file obtained from a relational database. Sometimes it is simply an itemized list in an HTML page. The attribute can be a text string that serves as an identifier for the entity in question, a text string that is the name of another entity and serves as a reference to it, a text string that characterizes the entity in some way but that is neither a name nor a reference, or, finally, a number or code that serves any one of these roles.

The result of the interpretation process is a text in a code language containing this information in the form of properties and corresponding values for a number of *entities*. An entity may be e.g. a country, a city, or a person, or in general, anything for which one can state a number of properties. The following is a simple example of two related entities in one of the two coding languages used by the CKL:

[7] `http://www.w3.org/2004/OWL/`
[8] `http://protege.stanford.edu/`
[9] `http://dbpedia.org/`
[10] `http://en.wikipedia.org/wiki/Main_Page, http://de.wikipedia.org/wiki/Hauptseite`

```
------------------------------------------------------------
-- Belgium

[: type country]
[: fullname "Belgium"]
[: iso-3166-code "BE"]
[: in-continent Europe]
[: official-languages {Dutch French German}]
[: shared-currency European_Currency_Union]
[: has-internet-domain "be"]
[: UN-member-date "1945-12-27"]

------------------------------------------------------------
-- Brussels

[: type city]
[: in-categories {is-capital}]
[: lang-name {[: Dutch "Brussel"]
    [: French "Bruxelles"]
    [: Swedish "Bryssel"]}]
[: in-country Belgium]
[: population "140,000"]
[: website "http://www.brussels-online.be"]
[: map-url "nationsonline.org/oneworld/map/google_map_Brussels.htm"]

------------------------------------------------------------
```

In this example, the property for the official languages of a country is a set of other entities, namely those representing languages. The curly brackets are used to enclose the members of such a set. The property called "lang-name" is an expression specifying the name of the city in question in a number of different languages. The mathematically oriented reader will recognize it as a so-called mapping. The property called "fullname" is used when the official name of e.g. a country is longer than the single word which is usually to name it. For example, the full name of the country of Iran is actually "Islamic Republic of Iran", and it may be shown in this property.

There are a number of steps between the original sources and a representation such as this. First of all, there is integration: the information for the various properties have come from different sources. In order to integrate, an identifier must first be selected. Sometimes the choice is obvious, but a policy is often needed to guide the choice, and sometimes individual choices have to be made in each case. For example, when a city has different names in different languages, which should be used as the identifier?

When an entity has several alternative names, different sources are likely to use different names for the same entity. Identification of the different name variants is therefore an essential part of the integration process. These differences may occur e.g. because the name variants from several languages are in common use (`Brussels` vs. `Bruxelles`, `Horatio` vs. `Horace`), or because there are longer and shorter forms (`Bosnia` vs. `Bosnia and Herzegovina`), or simply because several spelling variants are in common use even within the English language (`Argentina` vs. `Argentine`).

When the source information is organized as a table, ideally it is possible

to transfer the entries in that table one by one to properties of entities in the CKL system. However, there are many cases where this is not possible and the structure has to be changed. For example, in sources providing information about journals, there is of course one field for the name of the journal; however, this field often also contains ancillary information, and the journal name itself may have several logical parts. It may contain the same name in more than one language, it may contain both the full name and the acronym of the journal, it may contain the name of the publisher, the city of publication, or the year when the journal began publication or when it ceased, and finally there may be one part of the name that designates a set of related journals, and another part that designates the particular journal, for example as a specific 'series'. Some of this information may be necessary for identifying the journal: if several journals have the same name, then the name of the publisher or the city where it is published is important for unambiguous identification. On the other hand some information, such as the acronym, may be purely informative.

In such cases, the policy of the CKL is to attempt to separate the different parts of the title into different attributes. Thus there is one attribute for the name as such, another for the acronym, a third for the name (or identifier) of the publisher, and so on. This is because those attributes may be used for other purposes besides for discriminating between identically named journals, and it is useful to do the interpretation once, before publication, instead of leaving it to the user of the fact module.

Interpretation, as described above, can often be done on the basis of common sense, but sometimes it relies on the use of factual knowledge. For example, the following journal name occurred in one of our sources:

```
Economic Review - Federal Reserve Bank of Kansas City
```

It is easy to see that the second part of the name refers to the publisher, for a journal whose name would otherwise be ambiguous. However, another item in the same source is

```
Kansallis - Osake - Pankki. Economic Review
```

Here factual or linguistic knowledge is required to recognize that Kansallis-Osake-Pankki is the name of a bank in Finland and, thus, of the publisher. In this particular case the relevant information was known by the interpreter of the source, but a module for banks in the CKL would have been useful.

Another example of the use of background knowledge for interpretation of a knowledge source occurred when interpreting information from the Wikipedia concerning a certain Nikolai Korotkov, for whom the dbpedia reports that he was born in Kursk, Russia, but offering two different dates of birth, namely `1874-02-26` and `1874-02-13`. A likely explanation for this discrepancy is that they refer to the same day but using the Gregorian and the Julian calendar, respectively, since at that time Russia was still using the latter ([11]).

We foresee that there will be an increasing interdependence between facts and knowledge that are already in the library, and facts and knowledge that are interpreted in order to be added to the library, since already acquired facts may be useful in the interpretation of new sources.

[11] The difference between the two calendars was actually 12 days in the 19'th century, so there is still a difference of one day to be accounted for.

Already, some of the information in the CKL has been selected with this purpose in mind. For example, entities that represent countries contain one property for the country's top-level Internet domain, e.g. `uk` for the United Kingdom and `de` for Germany. This little piece of information is very useful for interpreting information about a new entity, namely, when the source provides its e-mail address or the URL (Internet address) for its website.

Some aspects of source interpretation, however, are based on knowledge of language rather than of facts. One important activity is the correction of spellings, names, and other relatively superficial aspects of the source information. Articles and other 'small' words in the names and titles are also often omitted or incorrect. It is surprising how many such mistakes there are in the sources that are found on the Internet. This is true even when the source text is in English, but the problem is far greater when names or other information are given in languages other than English. For the CKL we have chosen a European focus, and we aim to include names, whenever appropriate, in the main languages of the European Union and EFTA member countries, which in most cases means Germanic and Romance languages. Most of these use diacritical marks and umlauts, and these are often omitted in the sources; when they are included they are frequently incorrect.

In summary, the three main activities involved in interpretation of sources for the CKL are the assignment of identifiers, the identification and representation of information structure, and the correction of factual and linguisitic errors. However, we cannot guarantee that all factual errors are corrected, since we do not have the resources to check all details from the background sources for each individual entity. The information that we provide to users, after interpretation as just described, is essentially the information that was in the sources.

We do believe, however, that we significantly reduce the number of spelling and language errors, and that very few new such errors are introduced. Similarly, although the identification of structure often includes checking the background information, mistakes are possible. For example, we distinguish between different types of journal publishers, such as learned societies, academies, commercial publishers, etc. Our procedure determines that a publisher whose name contains the word "academy" will be classified as an academy in our sense, or as a learned society; however, if a commercial company chooses to use the word "academy" in its name, then it may not be correctly classified in the CKL module. An incremental process of feedback from users would be useful for solving this problem in the future.

In an ongoing project we are taking the concept of interpretation one step further and to the notion of *information analysis* of given data sources. This is the topic of a separate article ([12])

[12] `http://piex.publ.kth.se/reports/inca/002/` (persistent URL)
Erik Sandewall: Interpretation of University Namephrases. A Pilot Case of Corpus-Oriented Information Analysis. Project Memorandum, INKA project, KTH - Royal Institute of Technology, Stockholm, 2007.

Workdesk or Distributed Service?

This way of subjecting fact and knowledge modules to a form of publication encourages the use of a "workdesk" paradigm for the preparation and use of these modules. A spreadsheet software tool such as Excel is an example of a workdesk system: it allows the user to input information, to process existing information prepared by the same user, or by others, and also to define "stubs" or "macros" whereby routine aspects of the processing can be automated. The software system that we use in the interpretation of sources and for producing CKL modules has the same workdesk character, although it differs through its use of named entities and, more importantly, since the properties of entities can have a structure, as shown in the examples above. The software also contains extensive facilities for processing the notations that are used in web pages, in particular, HTML. We anticipate that other applications that use a similar workdesk paradigm will be able readily to import and use CKL modules.

The workdesk is a natural paradigm for the interpretation support system, since the existing factbase is used as a resource in the interpretation of additional sources, while at the same time these sources contribute to the factbase. This is in contrast to the distributed-service paradigm that was a part of the original proposal for the semantic web. In the distributed-service paradigm, the basic idea is that various websites set up services that can receive queries and return answers to those queries. Each service contains its own body of information, but it only delivers individual pieces of information to its software clients.

These two paradigms are not mutually exclusive, of course, since one can easily envision a workdesk type system that offers some distributed services, and that makes some use of distributed services elsewhere. By and large, however, the distributed-service paradigm seems to be particularly appropriate for situations where the body of information is very large and where it changes rapidly, since in these cases it is less attractive to use a copy of it that is not updated continuously. The workdesk paradigm, on the other hand, has advantages when the body of information is not enormous, and when it only changes occasionally.

In addition, the distributed-service paradigm may be advantageous for paid-for services, where the user pays a subscription fee, or where the software accessing the service makes a "micro-payment" for each access. On the other hand, the workdesk paradigm relies on the use of published fact and knowledge modules; this may make it more attractive for open-access resources, since these are typically rewarded by recognition by peers. Citation for the use of published fact and knowledge modules is similar to citation of conventional publications, and thus is easily understood; citation for the occasional use of an on-line service is less tangible, and may therefore offer less of a reward to the original author.

In addition to these considerations, there are technical issues to do with the reliability of access to distributed services and the performance of systems that are designed in that way. Those questions are however outside the scope of the present article.

Consequences for Libraries and Publishers

I have proposed that factbase and knowledgebase modules should be considered as publications, and that it is both appropriate and useful to treat such modules much in the same way as we treat traditional publications. If this approach gains acceptance, then it may impinge on libraries in two specific ways.

The first question is whether it will be relevant for a library to facilitate access to knowledgebase and factbase modules. In the short term the answer is probably no: these modules are freely available from the website, so why should there be a need for an intermediary? In the longer term, however, the answer may be different. In a scenario where there are many publishers of fact and knowledge modules, and a very large number of such modules are available, it will be a nontrivial problem to find one's way among these resources, and to make the best selections and combinations for given purposes. This task then becomes one for an information specialist, rather than for a software engineer, since it will require the ability to understand the character and qualities of the contents of available fact and knowledge modules.

The other question, which is of more immediate relevance, is whether the fact modules for the scientific publication domain that are currently available in the Common Knowledge Library can be of use for the development of software and services in the library. I hope and believe that this is the case, and invite the reader to visit the CKL website with this question in mind.

And what about publishers? This raises the question of what is the appropriate kind of organization for publishing fact and knowledge modules. The CKL is actually published in an organization within a university, namely, the KTH - the Royal Institute of Technology in Stockholm. Will it be reasonable, if and when the scheme scales up, to continue to publish such contributions through an organization within a university or affiliated with it (like a "university electronic press"), or will we see the development of larger, more centralized systems, for example under the auspices of learned and professional societies, or indeed commercial publishers? My personal belief is that we shall see all of these. However, I maintain that the most appropriate organization to develop, review and publish factbases and knowledgebases that contain information of central importance for a particular domain of science is, in fact, a learned society for the discipline in question.

Acknowledgements

This work has been performed while the author is a visiting professor at the KTH - Royal Institute of Technology in Stockholm.