Links to the Future

Harald Krottmaier

Institute for Information Processing and Computer Supported new Media (IICM)
Inffeldgasse 16c, A-8010 Graz, Austria
hkrott@iicm.edu

Abstract. *This article is about an upcoming feature in the Journal of Universal Computer Science ([J.UCS, 2002]) related on typed-links. We will introduce the concept of "Links to the Future" in context of the journal. Articles published in J.UCS are stored in an object-oriented database. Therefore features such as fulltext- and index-search are already available out-of-the-box. With the concept of "Links to the Future" it is possible to automatically generate a link to an already published article in J.UCS. Utilizing the paradigm of bidirectional links will help the reader to use this additional information about an article. It will be shown that this concept is limited to articles stored in the local database. However, a simple extension (i.e. working with so called surrogate-objects) will make it possible to enhance this idea to resources referenced in articles published in J.UCS. References to traditional resources (books and other printed material) are also analyzed. In the electronic environment several techniques (such as DOI, Digital Object Identifier) are available to refer to an unique electronic resource. Using surrogates of traditional material makes it possible to link to traditional material. Identification systems for traditional material (ISBN and ISSN) are already available and digital libraries should make use of them. To enable this feature it is necessary to add attributes to link-objects. Problems related to this upcoming feature are discussed. This article will give an outline and status report of the implementation work.*

Keywords: Content Link; Reference Linkages; DOI; ISBN; ISSN

Received 21 Nov.2002; Accepted 4 Jan. 2003

1. Introduction

The Journal of Universal Computer Science (J.UCS) is now running for more than 8 years. It was one of the first electronic published journals who implemented features such as personal- and public-annotations, multi-format publications, multi-categorization, etc. However, readers from high-quality electronic journals are used to use more highly sophisticated features, such as automatic reference analysis, similarity search between documents and other features using knowledge management technology. In this article we will give an overview of an upcoming feature in the research environment related to linking technology using typed-linking.

Since the journal is based on an object-oriented database it is possible to add arbitrary attributes to objects stored in that database, i.e. it is possible to add attributes to links. The features described in the following are based on this technology and uses linking technology such as typed- and bidirectional links.

2. About the Feature

The idea of "Links to the Future" is to automatically create links *from* previously published and related material *to* the new published material. This task is difficult in the global world, but restricting the scope to a local environment will make the task manageable. It is obvious that bidirectional links are of great advantage in this application. In our environment we are going to create links *from* the newly published paper *to* related material stored on the same server. Using bidirectional linking technology will make it possible to display links also at the target of the links, i.e. the referenced article.

Several questions must be asked in this context:

1. How to identify related material?
2. How to find related material?
3. How to overcome limitations?

In the following paragraphs we are going to answer these questions. The first question can easily be answered in a restricted environment. In the area of scientific publishing with well known structure (i.e. abstract, content, special reference section) related material is listed in the reference section. According to the given style-guide of the documents the reference section is usually located at the end of each article and it starts with the keyword 'References'. Each entry starts at a new line and begins with something like `[Name(s), Year]`. Technically it is therefore not a problem to parse all reference-entries and store them in an easy to process way. Unfortunately it is difficult to identify names of the authors, title of the publication etc. If the authors of a document consider the style-guide it should be possible to extract the information without too many mistakes.

To be as general as possible we accept PostScript and PDF-Versions of articles. A conversion to an easier to process format (at the moment we use simple ASCII-encoding without any markups…) must be performed by tools (`pdftotext`, etc.). Experiments currently performed at a sample-set of publications discovered problems with these tools. We recognized that authors do not follow style guides as severe as they should. Therefore several attempts must be performed to find the right information.

Each entry in the reference section is stored separately in a special object. The plain-text of the reference-entry without any markup is stored as text-object as well as attributes (such as title of publication, name of authors etc.) are added to the specialized text-object. Since we are currently facing problems with adding metadata to single parts of an entry we limit extraction of information to the name of authors and title of the reference. If there are additional URLs stored in the text-part of the entry, we save this value as well.

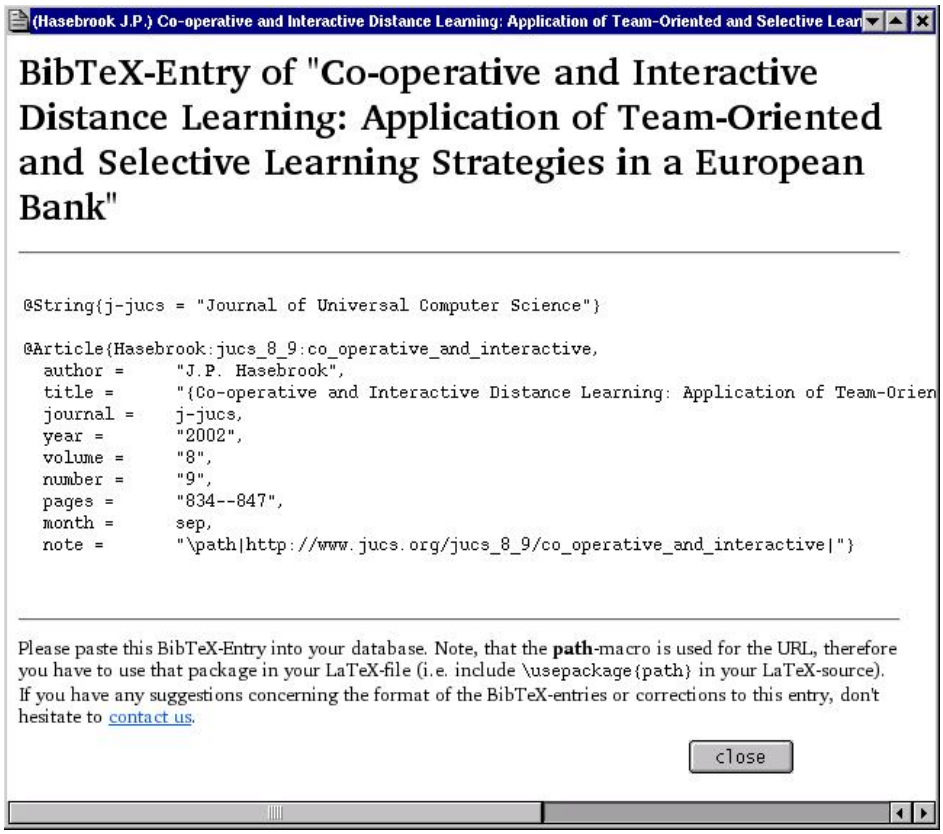

Fig. 1. BibTeX-Entry for an article

The second question is about finding the right resources. This is difficult because of the low quality of metadata we are able to extract out of the plain-text of a reference-entry. In the first prototype of the "Links to the Future"-feature in J.UCS we limit links to articles published in J.UCS. Therefore we explore whether the entry is related to an J.UCS-article or not. This is performed via regular expressions ([Friedl, 1997]) of the name of the journal and of the URL-prefix of J.UCS.

To make life easier for our programmers we offer well structured and *correct* entries for each article published in J.UCS (see figure 1). If the reference-entry is created with the automatically generated BibTeX-entry we can be sure that the entry is correct and the URL is given in the reference-entry. Even if the URL is not given in the reference-entry we are able to generate the URL out of the Title of the article and the related information about the volume- and issue-number. Since we offer different formats of the published content, we have to normalize the URL according to the "entry-page" of each article.

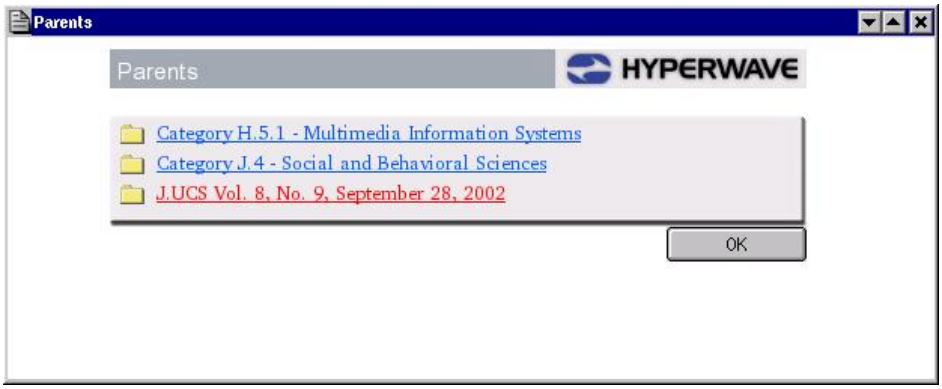

Fig. 2. Dialog displaying parents of an article

Once the URL of the referenced article is known (and the URL is valid), the bidirectional link is created and special attributes are added to the link-object. "Links to the Future" are displayed dynamically via server-side scripting methods. A very similar dialog is shown in figure 2. In the current environment links to so called "Parents" are implemented via standard Hyperwave visualization methods. The object-oriented database used in the environment supports parent-children relations in a non-circular way. If the new published article is inserted or prepared with specialized tools, these parent-children relations are created.

2.1 Limitations

Several limitations are inherent in the described approach. However, these limitations may be removed by enhancing the current implementation. In the following we are going to describe the obvious limitations and propose a solution to them.

Links to correct cited references: At the moment it is possible to automatically link to correctly written references. However, several authors make mistakes when writing reference-entries. Modern knowledge management tools such as similarity-search allow a

moderate error-correction. Restrictions of links to intra-server links will improve the situation. One can imagine that given the correct name of an author and a mistyped title of an article will lead to the correct article even without using modern knowledge management tools. However, if both of them are mistyped, knowledge management tools will help in finding the right publication.

Links *to* published articles in J.UCS: As the procedure described above had shown, bidirectional links are inserted and stored in the database. Since these links must also point to objects stored in the database we have restricted targets of links to intra-server objects. However, one can create surrogate-objects in the database representing any type of reference. In the used database system it is possible to create so called "remote-objects" with special attributes. These surrogate-objects will be used as link-targets, therefore bidirectional links may be created. When extending the current prototype attention must be payed to duplicated objects representing the same remote reference. This might be tricky because content is duplicated and accessible via different URLs. In this case similarity algorithm will help in finding the duplicated contents.

Links *from* published articles in J.UCS: Once articles are published in J.UCS it is possible, to extract needed information out of the parent/children relation or to get information out of the system-immanent link-database. The number of articles in J.UCS is limited and it must be possible, to add a list of 'incoming'-links to any article. The goal is to add this feature to any electronically available cited articles. Therefore a surrogate-object as described in the previous paragraph will be created for each electronic-available reference. Using sophisticated parsing techniques (as implemented e.g. in the Google-search engine, [Google, 2002]) it is possible to search for `a`-tags via special search queries. Obviously only indexed pages are search-able. However, many publishers such as ACM make material available for indexing only.

Duplicates of Resources: It is very easy to copy electronic resources. Unfortunately it is not very easy to detect duplicates. It is very common to link to the article at publisher's server and not to local copies of an article. Currently another lookup service is very popular: the concept of "Digital Object Identifier" (DOI, [DOI, 2002]). DOI implements "stable identifiers for electronic resources". However, it is necessary to detect duplicated entries of resources. Knowledgemanagement tools such as similarity search between documents are already available and must be used.

As one can see, the given limitations of the prototype are easily removed by extending the implementation.

2.2 Links to Traditional Material

In large digital journals and electronic publishing systems articles are already parsed and links *from* the reference in the text *to* the reference-section are automatically generated. These links are called "intra-document links".

However, it is often desirable to explore the reference itself. Electronically available references may be linked as described in the previous section. Traditional material must be

handled in a different way. Depending on personal preferences these links should be either directed to book-resellers — such as Amazon — or directed to the catalog of local libraries. Currently we are implementing such a linking feature to an electronic catalog available in Graz, University of Technology.

The key issue in linking to traditional material is the unique identification of the resource. The following identification systems are currently available and should be used for identifying traditional resources:

International Standard Book Number (ISBN): [ISBN, 2003]: "The ISBN is a unique machine-readable identification number, which marks any book unmistakably. This number is defined in ISO Standard 2108. The number has been in use now for 30 years and has revolutionized the international book-trade. 165 countries and territories are officially ISBN members. The ISBN accompanies a publication from its production onwards."

International Standard Serial Number (ISSN): [ISSN, 2003]: "The ISSN is an eight-digit number which identifies periodical publications as such, including electronic serials. More than one million ISSN numbers have so far been assigned. It is managed by a world wide network of 75 National Centers coordinated by an International Center based in Paris, backed by UNESCO and the French Government. The ISSN is used by various partners throughout the information chain: libraries, subscription agents, researchers, information scientists, newsagents (through its barcode version)."

With the appropriate ISBN for books or ISSN for periodicals it is possible to identify the respective entry in an online-catalog of books. Unfortunately these two tags are not available in every reference entry. In reference management systems (such as BibTeX) this attribute is *not* mandatory.

In the first implementation we will concentrate on reference-entries with given ISBN/ISSN. Thereafter we will try to explore the entity via the title and author.

3. Conclusions

Links from the reference section to the appropriate resource are necessary and users of electronic-publishing facilities are used to it. This article opened the discussion for further features related to reference linking: The idea of "Links to the Future" is the reverse idea of "Links to References". This feature links the reference to the resources. Therefore it is possible to get all articles, which refer to a reference. In the current implementation we are going to restrict links to intra-server documents. A future refinement will expand the idea to other resources as well.
Links to traditional material such as printed books and periodicals will enhance usability of the reference section of an article published in J.UCS. On account of personal preferences links will either be directed to a bookseller or to the local library. As prove of concept we are going to use a locally available catalog of the library.

References

DOI (2002). Digital Object Identifier. http://www.doi.org.
Friedl, J. (1997). Mastering Regular Expressions. O'Reilly Associates, Inc.
Google (2002). http://www.google.com .
ISBN (2003). International standard book number. http://www.isbn-international.org .
ISSN (2003). International standard serial number. http://www.issn.org .
J.UCS (2002). Journal of Universal Computer Science. http://www.jucs.org .