Understanding and Supporting Volunteer Contributors: 
The Case of Metadata and Document Servers

Bettina Berendt

Humboldt University Berlin, Institute of Information Systems 
Spandauer Str. 1, D-10178 Berlin, Germany, http://www.wiwi.hu-berlin.de/~berendt

Abstract

The semantic interpretation of text remains a hard AI problem, and it becomes particularly relevant when large archives have to be built and searched. To aid in this process, advanced Web-based publication services (that is, document servers with surrounding infrastructure) adapt methods from traditional archiving, employing bibliographic metadata to improve data quality. However, most services cannot afford to pay for ex-post annotation, so the document authors themselves must provide metadata content and markup—additional work that may deter them. Services therefore have to acknowledge that while there may be many authors who are potential contributors, the steps to actually publishing with the service and to delivering high-quality metadata are voluntary; i.e. authors who decide to do so are volunteer contributors. In order to win contributors, a service must therefore understand its potential contributors’ concerns, and it must evaluate its capabilities of addressing them. We present a case study of a large university document and publication server. Surveys and Web usage mining identified which kinds of knowledge can or cannot easily be gathered from volunteer contributors. We also describe a tool that aims at improving the HCI incentives by employing text mining methods and presenting an easy-to-use interface to ensure correct markup. We expect that the recommendations and technology and interface concepts can be generalised to the needs of a range of other volunteer services with similar incentive structures.

1 Introduction

Much of today’s Web publication activity still relies on making text collections electronically available and searchable in the ways afforded by currently popular formats and software such as PDF. However, this approach has a number of limitations. Technologically, these formats are often proprietary, and it is far from certain that today’s files will still be readable in 50 years’ time. Semantically, they are restricted to full-text search and whatever information external search engines can extract from the syntax of the document itself and the Web as a whole.

Therefore, in order to provide the advantages of full-blown archiving—long-term readability, unique addresses as identifiers, authenticity, and semantic accessibility—advanced document servers rely on open syntactic standards such as XML, and on metadata such as those proposed by the OAI (www.openarchives.org). To ensure high-quality documents, most document servers employ a (usually small) paid staff who test the documents’ and the metadata’s suitability and quality, correct obvious errors, convert Word documents into XML, etc.

However, only the authors have the knowledge to provide the metadata content; and usually, they must also perform a large proportion of the technical (meta)data preparation. This requires additional work and additional qualifications, and it influences an author’s decision whether to become a contributor or whether to revert to lower-standard online and/or regular offline publication. Thus, authors who do contribute are volunteer contributors; in order to attract them, a service must understand potential and actual authors’ concerns and wishes.

The first purpose of this paper is to present a case study that exemplifies a methodology for investigating what kinds of knowledge can and cannot be gathered from volunteer contributors. This comprises AI methods from Web usage mining and knowledge representation.

The second purpose is to propose hypotheses and actions based on our results that address the question whether useful knowledge collection can be successfully deployed as a side effect to user’s activities, e.g. searching or using the Web. This includes using AI methods to illustrate principles of successful cooperation to potential volunteers.

The third purpose is to propose that a successful strategy for winning contributors, or design principles of interfaces friendly enough for beginners and advanced users to use, should be based on a wide notion of “service interface” comprising four types of incentives: motivational, institutional (authors are forced to supply content, or to supply it in a certain way), and computational. The computational incentives include HCI incentives (interactivity to make the creation of content/metadata more usable) and machine learning incentives (tools that extract metadata proposals). Here, AI methods from text mining are used to aid the users in their task of generating metadata.
2 On the ease of gathering different kinds of knowledge from volunteer contributors

The aim of this case study was to evaluate the outreach towards potential and actual authors of the Document and Publication Server at Humboldt University Berlin (http://edoc.hu-berlin.de), i.e. to understand how the HU EDOC service reaches its potential or actual volunteer contributors. In order to carry out an evaluation of a Web site and its accompanying offline services as multiple user-contact channels, it is necessary to know the service’s goals and target groups, as well as their attitudes, wishes, and behaviours, see (Berendt et al. 2003; Berendt et al. 2004).

2.1 EDOC – server and service

EDOC has been operated as a document server for the publication of HU scientific work since 1997. Doctoral dissertations have so far accounted for the largest proportion of published works. Electronic publication is attractive for PhD students because in Germany, doctoral dissertations must be published to obtain the degree, and publication in a publishing house is time-intensive and costly (from $1000 to $5000 or higher), but free on EDOC.

To publish on EDOC, authors must use the provided templates and structure their dissertations accordingly. The advantage is twofold: First, properly formatted texts can be converted, by EDOC staff, into DiML (the SGML/XML-based Dissertation Markup Language) to ensure long-term archiving, and second, the markup contains semantics such as text structure and citation elements that support structured search. Authors are offered intensive support: information materials, training courses on the use of the Word template, and telephone, email, or in-person support.

The EDOC service has two main target groups: authors and readers of academic documents, in particular, doctoral and “Habilitation” dissertations (German post-doctoral degree). The objective is to convince authors of the advantages of electronic publishing and encourage them to create content, and to encourage readers to access and utilise this content (see also Wendland, 2002).

2.2 (Potential) contributors—questionnaire studies

One major question of the service team was why the majority of potential contributors do not take advantage of the opportunity to publish online. Since the beginning of the EDOC projects, a relatively stable (and arguably small) percentage of dissertations has been published online: around 20% of all HU dissertations (13% if the medical faculty is included in the statistics). How should this proportion be interpreted? Does it indicate that the remaining doctoral students and doctors do not wish to publish online, that they do not feel capable of publishing online and/or adequately supported in their attempts, or that they are simply unaware of this possibility?

Two online surveys were developed, covering the non-medical faculties and the medical faculty, respectively (for details, see Berendt et al. 2003, Wu 2004). Emails were sent to lists comprising nearly every member of the target group “authors of doctoral or Habilitation dissertations” (1180 / 1290 recipients in the first / second study). Response rates were 13.7% and 12%. Main findings were:

The majority of the target group is willing to contribute their dissertations as long as it does not diminish their reputation. Book publication in a publishing house remains the most preferred option among the non-medical faculties (45.9%), the main reason being that “in my discipline, online publication is considered informal and not as highly valued as a real publication”. 19.8% intend to publish online only; an additional 62.2% favour online in addition to book publication. Reasons include cost, speed, and access. In the medical faculty, total online preference is similar (86%), with 32% intending to publish online only.

So while many authors are willing and probably able to contribute their “raw data” (= their dissertations), they are often not aware of the associated need to also develop and contribute “metadata” (= the dissertation template markup). However, the contribution of these metadata is a necessary condition for publication and for the long-term usefulness of the raw data. Unfortunately, results on the use of templates and metadata were very discouraging:

Authors take concrete steps towards digital publication late (and are thus poorly informed about requirements of the contribution process).

Metadata generation/formulation is considered to be one of the major problems. Templates are not used by the majority of authors, and the use of templates for scientific writing is not discussed in courses (asserted by 76% and 94%, respectively, of the respondents in the second study).

A major bottleneck in contributor acquisition is information about the very possibility of being a contributor. One of the most striking findings was that when asked how they found out about the opportunity of publishing their dissertation on EDOC, nearly half of the participants answered “from this questionnaire” (44.1% / 28%). 28.3% / 42% wished they had learned earlier about the digital publishing opportunity.

2.3 Readers and contributors—Web usage mining

To complement the authors’ self-reports, we analysed user behaviour in the field. This allowed us to also study the behaviour of the target group “readers” (for a supporting survey study, see Berendt et al. 2003). In general, every contributor can also be a reader, and many readers can also be contributors. So are there patterns that indicate that the site turns readers into contributors?

From EDOC’s Web server log, we extracted a sample of 10,992 sessions (210,655 requests) from one week in 2003 (near the end of the first survey). Data analysis comprised semantic enrichment, clustering, classification, and association rule and sequence mining. Main findings were:

Doctoral dissertations are the most-often requested type of content.

Readers and contributors are distinct groups; readers are not led to accessing contributor content (at least not in a single session).
Only few people use the internal search engine, and they do not experience structured search as an effective or efficient search option. The use of external search engines makes access to dissertation full-texts more likely. Possibly as a side-effect of this direct route into full text, 36% of full-text readers did not read any abstract and can thus be said to not have obtained a meta-level view of the text they accessed.

3 Deploying knowledge collection as a side effect to users’ “normal” activities

Summarising the findings, we can state that in order to increase volunteer contribution volume, services should meet a number of marketing challenges. First, the site’s surrounding infrastructure (in our case, “multipliers” such as professors or libraries) should improve information policies, and respect specific bottlenecks (in our case: the lack of a single point of contact to dissertation authors). Second, the site itself should attempt to convert readers into contributors by an interface for readers that de-emphasises the static document-collection aspect and emphasises the dynamic knowledge-repository aspect.

However, authors need not only contribute their dissertations at all, they also need to contribute structure and format through metadata. We observed that authors do not perform structured writing, and that readers do not use structured search. This points to an education challenge. Why should authors embark on the laborious path of structured writing when they never use such structure; conversely, it is necessary to use structure in writing in order to profit from it in search. Therefore, the conditions under which this type of text is generated (or not) become important. Lecturers can encourage the practice of structured writing by setting an example and by making the use of templates and markup a course requirement.

This should be motivated rather than enforced: We have observed that many students have extensive problems with selecting and citing literature, and that one of the reasons is that the functions of structured citation are largely unknown (what does the publication location tell us? Why is it important to supply not only a URL, but also the original book, journal, etc. source?). In teaching, I focus on the information contained in this structure, and use Internet metaphors to explain citation networks: the AI method of link analysis that students experience as useful when they employ Google, as an illustration of principles of successful cooperation (see Berendt 2003). The aim of this approach is to engender a sense of belonging to and trust in the academic community as well as the experience of competence—necessary prerequisites of “pro-social motivation”, which along with “fun” have been identified as key intrinsic incentives in studies of the success of volunteer contributions in the open-source movement (Osterloh, Rota, & Kuster 2003).

4 A tool to help volunteers create semantics

In spite of the influence of social factors, a large proportion of content generation remains solitary work with software tools that often leave the author with little sense of competence or fun. To support the experience of competence, we are developing authoring tools. We first focused on bibliography markup, which was described by many authors as one of the most time-consuming, boring, and error-prone activities in the conversion of a previously unstructured dissertation into one with EDOC markup.

Today, the typical author contributing to EDOC uses Microsoft Word, so she has to mark a text passage (say, an article’s title) with the mouse and then choose the markup element from a dropdown menu. This often leads to errors like the following, which may go unnoticed even in the careful post-processing exercised by EDOC staff (passage taken from the DiML version of an EDOC dissertation):

```
<BIBLIOGRAPHY><FLOAT><PAGENUMBER>136</PAGENUMBER></FLOAT>
<HEAD>Literaturverzeichnis</HEAD>
<CITATION WORKTYPE="journal" PUBLISHED="PUBLISHED">
  <CUT ID="bib-15">[1]</CUT>
  <ARTICLETITLE>Ultrafast energy transfer in LHC-II revealed by three-pulse photon echo peak shift measurements</ARTICLETITLE>, <WORKTITLE>J. Phys. Chem. B</WORKTITLE>, <PUBDATE>2000</PUBDATE>, <NUMBER>104</NUMBER>, <PAGES>2908</PAGES>, ...
</CITATION>
```

Errors in citations are a hard problem because citation styles are not uniform, even within a discipline, and seldom used consistently, even within a document. These errors are an important problem because they may have serious consequences on citation indexes, authors’ reputation, and the memory of science as a whole (Cardona & Marx 2004).

The Intelligent Literature Markup tool is based on information extraction using regular expressions (Grover et
al. 2000). Its system architecture is shown in Figure 1. The prototype was implemented on a Windows PC, in which the MS Word environment represents the author’s PC and the text mining functionality is realised in a VMware environment under Linux. Since it is unrealistic to expect non-expert users to run this system configuration (or even to download and install additional software), we propose an implementation as a Web service on EDOC itself. This makes secure data transport essential (currently, via SFTP).

Figure 2: User dialogue: journal articles and monographs.

For the user, details of this architecture are invisible: She marks a complete, unformatted reference with the mouse, and is presented with a proposed translation into the reference’s elements. As Figure 2 demonstrates, the user is then able to accept the title (etc.) shown in the text field, or to correct it if the tool has assigned text portions erroneously. The user is also able to choose the citation style via radio buttons, thus informing the system how to interpret the reference text marked with the mouse. The end result is translated into DiML. All processing and communication with the text mining software are handled by a Visual Basic for Applications (VBA) macro.

The distributed architecture has the further advantage that all system updates and extensions can easily be implemented and made available to all users, and that complete and current records of usage are available and can be analysed in order to evaluate the system (anonymisation methods will be used to ensure privacy and user trust).

5 Summary and conclusions

We have presented a case study of a system that relies on attracting volunteer contributors for its contents, and described empirical studies of its (potential and actual) users’ expectations, opinions, and behaviour, identifying lack of information and structured-writing skills as the two major bottlenecks. We have sketched educational approaches to teaching structured writing as a strategy to overcome these bottlenecks that focuses on the social aspects of producing high-quality documents, and described a software tool as a strategy that focuses on the individual aspects of this process. We believe many other volunteer services share much of our case study’s incentive structure, and therefore expect that the technology and interface concepts can be generalised to their needs.

Future work will first aim at an evaluation of actions based on the findings described here. This evaluation will start with user tests of the Intelligent Literature Markup tool in controlled laboratory conditions, investigating to what extent authors regard it as a helpful cognitive tool. Planned extensions of the tool include further citation styles and learning mechanisms. Testing whether the use of the tool in the field will improve contribution volume and quality is a methodological challenge because the introduction of (any) new service feature will be confounded with the ongoing developments in students’ acceptance of digital publishing. We plan to evaluate this by seminar group studies.

References


I thank Thomas Dammeier and Thomas Posner for developing the Intelligent Literature Markup prototype.