PALESTRA

WHY NOT LET THE COMPUTER SAVE YOU TIME BY READING THE TAXONOMIC PAPERS FOR YOU?

D. Agosti[1], C. Klingenberg[2], G. Sautter[3], N. Johnson[4], C. Stephenson[5], T. Catapano[6]

[1]Plazi.org; Dalmaziquai 45, 3005 Bern, Switzerland. E-mail: agosti@amnh.org

Introduction

The traditional role of systematic publications, or more exclusively the description of new taxa, has been to communicate to the word that new taxa have been discovered, to provide guides to its identification and to fulfill the criteria of the International Code of Zoological Nomenclature (ICZN, 1996) to make the names available to the scientific community. Until about 1996, when the Internet became established and the first digital publications appeared, all the descriptions have been printed in journals, books or adequate materials, such as Bill Brown's Card Index (BROWN, 1964). Despite the Internet, a minimal set of hard copies of digital publications are needed to fulfill the ICZN requirements, and a change allowing digital publications only is in preparation. Zootaxa's explosion in the production of new species is the best indicator of the taxonomic community in making usage of digital media (KNAPP et al., 2007).

Though the use of digital copies of publications has its advantages over paper copies, such as the ability to send them around by email, or even better allow open access and thus allow anybody indiscriminately access to it, the real advantage of the Internet has not been capitalized, and in fact, the behavior of the readers is still the same as in the Gutenbergian age of paper copies, where one had to read each page.

The real value of the Internet is though that the machine can take over much of the reading, searching and data mining. For example, in the collection of PubMED, probably the largest collection of abstracts on biomedical literature, 16 Million abstracts are available. In an experiment by the Wilbank at the Neurocommons, 400,000 abstracts have been isolated and about 100,000 have been analyzed. This resulted in over 30,000 relationships between 5,500 genes and proteins (WILBANKS, 2007).

Though taxonomic publications are only a minor subset of the biomedical publications, for example for ants ca 2,800 describing all the species only (AGOSTI& JOHNSON, 200??), this publications are very dense in information. The descriptions are highly formalized, and increasingly in modern publication include an extensive morphological description, often very detailed collecting events in the materials examined, plus some summary data the distribution pattern, biological relationships, ecology and behavior and the taxonomic relationship, and images.

At the moment, this data can only be harvested one publication by the next. Clearly, in the future, most the data will originate from dedicated databases, such as character vs. taxon matrices, distribution databases, image databases. But in the meantime, and until everybody would have access to such databases, we envision the intermediate solution to provide a formal framework on how to model and transfer old and current publications into a machine readable form. The technical set up will be discussed below.

Access to scientific publications calls for the clearance of the respective copyright licenses. This is a very important issue and various solutions are proposed, such as the Green and Gold road to Open Access (HARNARD, 2007), whereby the former calls for the self-archiving of all your research publications, something which is increasingly possible. Each individual scientist can furthermore negotiate with the publishers what sort of license are being used for a scientific publication, which should result in publications for which the right is not completely transferred to the publishers, but kept for further useage by the author.

The best way to the future will be though to demonstrate the power of access by building up a stock of machine readable publications and run analysis which are not possible without massive amounts of data and thus access to the respective data. For example, a global analysis of ant diversity at species level would only be possible by mining a representative part of the printed record. This is much faster then to go through all the collections and identify all the specimens, which is not doubt even more powerful.

Access to the printed record will also allow serving its content, so that specific Web sited could be built on

[2]Staatliches Museum für Naturkunde Karlsruhe, Abt. Entomologie, Karksruhe, Germany.
[3]Universität Karlsruhe (TH), Department of Computer Science, Karlsruhe, Germany.
[4]Ohio State University, Insect Collections, Columbus, Ohio, USA.
[5]American Museum of Natural History, Libraries, New York, USA.
[6]plazi.org, Columbia University, New York, USA.

the fly, which then, in return would allow to add additional data, such as those collected in field surveys or biological studies. In return, such pages, like EDIT's scratchpads, would allow amalgamating data from various sources and supporting taxonomic studies to revise a particular taxonomic group often used in various non-taxonomic studies. Access to the printed record will also allow to formulate ecological hypothesis, such as which ant is interacting with which organism, or more trivially, where particular specimens can be found, both in collection and the field.

Though it is not yet clear how much it will cost to make all the publications machine readable, this process can be seen as a one-off complete removal of the taxonomic impediment, that is providing access to published taxonomic literature.

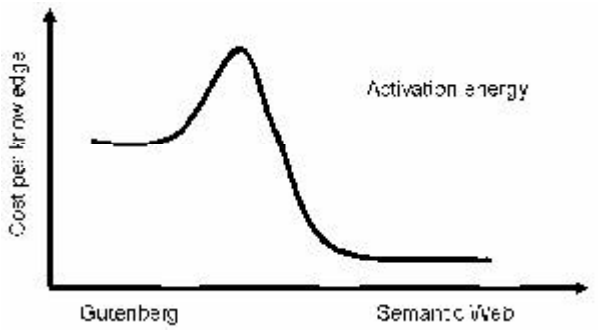

Fig.1 - Data conversion seen as seen as an activation energy from a current state of knowledge to a one in a semantic Web environment.

Since this conversion comes at a cost which is currently not covered by science projects and the conversion depends on specialists input which at the same time provides increasing support for scientific studies, we imagine, that the respective taxonomic communities might be the ideal actors in this data conversion process.

In the reminder of this contribution we will discuss plazi.org, a dedicated web site and project to allow taxonomic communities to support this conversion process and to harvest this newly developed data repository.

This process is as much a data conversion process as much as an experiment on how this can be done efficiently. The results from this project might play a crucial tole in the advent of the launch of Zoobank, a global registry for zoological names, which still is missing huge junks of published names, or the even more ambitious Biodiversity Heritage Library, aiming at scanning in a first round all the holdings of taxonomic literature in some of the large Anglo-Saxon natural history institutions (BHL, 2007)

The mechanics of access

Plazi.org, an independent Web site dedicate to provide access to taxonomic literature by looking into removing existing barriers, such as copyright, finding technical solutions and create the necessary environment so that the conversion will take place. It is built on Drupal Content Management software and DSpace, a dedicated software for digital literature repositories. The latter has an important function to help to find digital copies of the digitized copies, irrespective of where they are hosted. Publishers use increasingly Digital Object Identifyers (DOI). Handles are a cheaper solution used by self archives of scientific literature. They both work on the principle that the respective digital archives are registered with a resolving mechanism, which allows to resolve a given handle:

http://hdl.handle.net/10199/1379 will resolve the actual physical site http://plazi.org:8080/dspace/handle/10199/1379, because currently, the pdf version of this publication is hosted here. Google and other, more specific search engines, will the always be able to find the respective pdf or additional digital versions of the document.

Machine readability means, that the machine understands what the bitstream it is looking at means. Schemas are one way to communicate with machines. They generally model the logic content of the bitstream or document in this case. It first defines that it is a publication and then what is included, such as names, geographic names, description of species. TaxonX is such a dedicated XML schema, which in itself is complementing other now widely used schemas, such as Darwin Core for specimen data and names.

The mark-up process, that is adding in the appropriate places the elements, is in itself a candidate for automation. But successful mark-up depends on the best possible, ideally error free text recognition of the originally printed text. This again depends on ideal scanning of the print, whose decisive elements are the selection of the original, the contrast / brightness optimization, or the resolution of the scan.

GoldenGATE is our dedicated editor, which is essentially a framework for which particular tools can be built to resolve specific steps, such as finding taxonomic names, completing taxonomic names, interacting with dedicated name servers, such as Zoobank and retrieving globally unique identifiers (GUIDs, ie. Life Science Identifyers in ZooBank) or getting handles for the literature in the bibliographic references. Especially older publications are in most cases easily understandable by a person, they pose a rather great challenge to automation. Avoiding adding artifacts in this process, GoldenGATE has two additional features: it can be trained by building

analyzers for particular kind of publications or taxonomic group which can be shared among the users of GoldenGATE, and it has an interactive mode, which at least presents unclear results for manual resolution to the editor.

GoldenGATE is run on the client side to allow working independently of Webaccess, and only at moments to add GUIDs. or the finished document.

Plazi.org offers the entire document management tools from entering the metadata of a particular publication to upload and download documents at various predefined stages of document conversion (Community Mark-Up Server) from pdf, to OCR-ed document to marked-up TaxonX-documents, to the storage, indexing and Webservices and search-tools to the marked-up elements of the processed documents (Fig 2).

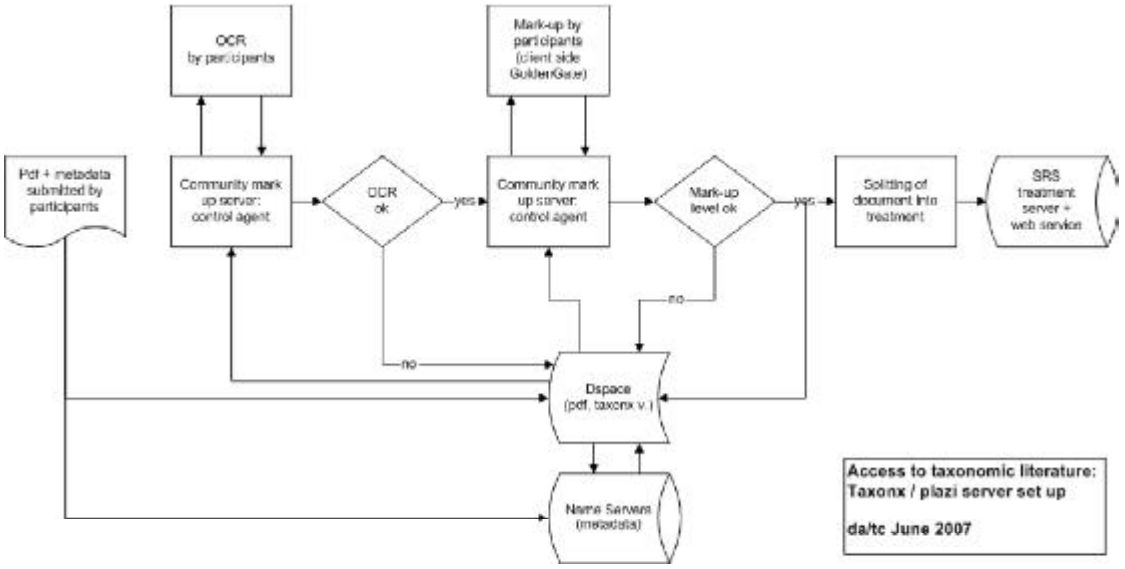

Fig.1 - Data conversion seen as seen as an activation energy from a current state of knowledge to a one in a semantic Web environment.

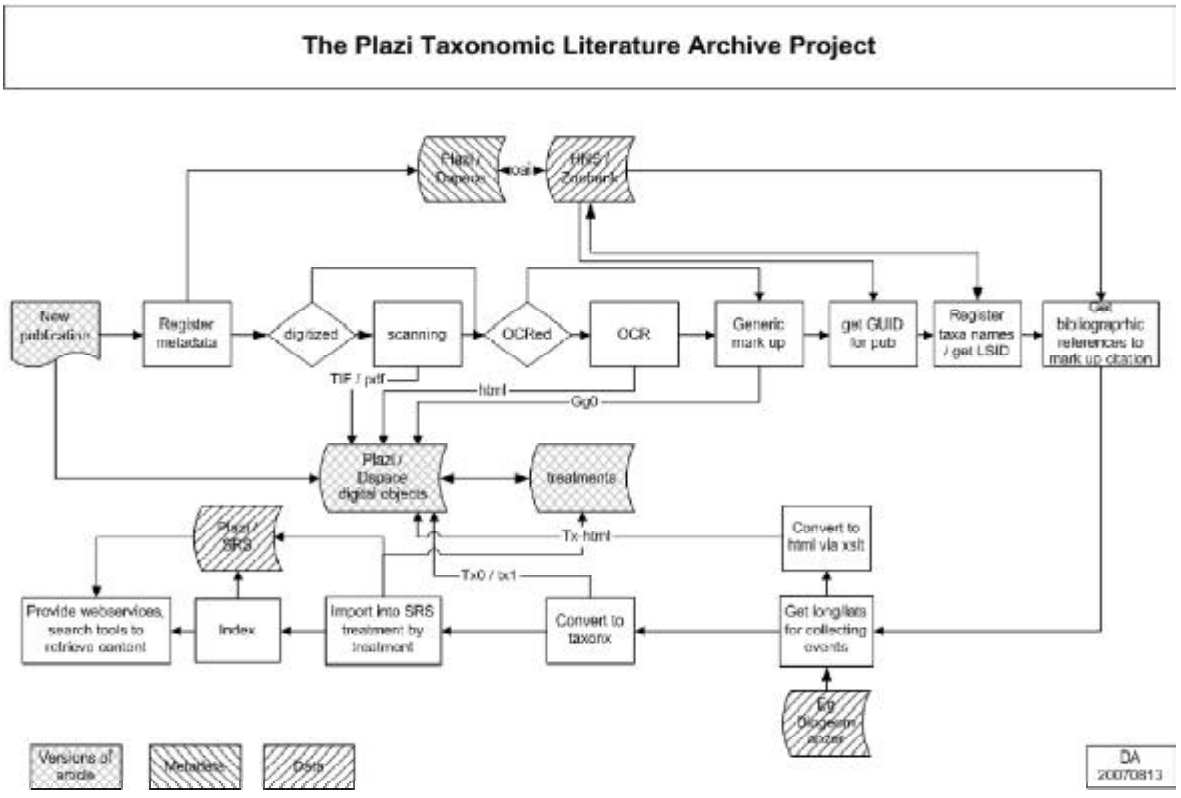

Fig 3 - Detailed data flow in Plazi, from a publication to specific holdings of data and respective metadata.

The real values added to the documents are not only the mark up per se, but the normalization of the document and that that all the relevant marked-up documents are given the respective GUIDs either from an external service, or internally for when no such service exists like for treatments. (Fig. 3).

Normalization means, that the content is not just marked up, but a standard version of its content is provided.

Temnothorax korbi would then read like this, including the GUID referring to the unique number of this taxon in the Hymenoptera Name Server.

```
<tax:name>
<tax:xid source=”HNS” identifier=”183325"/>
<tax:xmldata>
        <dc:Genus>Temnothorax</dc:Genus>
        <dc:Species>korbi</dc:Species>
</tax:xmldata>
T. korbi (Emery, 1922)
</tax:name>
```

With the respective registration of the GoldenGate treatment server, the server which allows direct or through an advanced search function access to and retrieve treatments (SRS), with services like TAPIR and Zoobank, the computer can begin crosswalks from one data element to another. It could for example be shown, where has which specimen be published or referred to? Which publication has been cited how often? But it allows also any kind of new application, such as what species do I find where.

Since the mark-up process can be stopped at various degrees of granularity, from a mark-up including just names and treatments down to individual characters and details in specimen records like collection numbers, etc., it allows a continuous refinement of the access depending on the requirement of the users.

The goal of Plazi.org is to provide access to data in publication. Thus the emphasis is given to provide data aggregators such as the Encyclopedia of Life (EOL, 2007), or data creators EDIT's Scratchpads (NEOTROPICAL ANTS (FORMICIDAE, 2007)) and dedicated Webpages such as the planned pages for the Neotropical Catalogue of Ants, antweb.org, the German Ant Types Pages (www.anttypes.org), or the Hymenoptera Name Server Webservices and other tools to access the database.

A human readable interface is provided for specific searches for any elements stored in the database, and basic facilities are offered to query and visualize its content such as plotting the collecting events on a google map.

Plazi is still as much as research project as it is increasingly a production site. It houses currently collections of ant literature and those for lice and selected spider taxa are being added. This comes at a cost, that certain features are changing, but at the same time that changes can be accommodated. The mark-up process is slowly approaching a minimal time per page for an experienced editor, which clearly needs be a person with some minimal understanding of taxonomy and the respective taxa.

Conclusion

That computers can read and analyze our systematic publications is rapidly becoming reality. Plazi.org is one of the first integrated system. As a pilot system, its development fosters not only the development of novel tools to read and understand publications (Name searching algorithms like FAT (SAUTTER et al, 2006), GoldenGATE editor), but shows the enormous amount of work needed to convert a very heterogenous body of published work. Clearly, it could be argued, that part of the time gained of having our publications marked up and available in dedicated servers would be worth a collaboration with the users by spending a fraction of this gained time in the conversion process.

More importantly, the expertise and development in the workflow of taxonomists point out, that we need in future to add the mark-up during the preparation of the manuscript, and most likely make the publications as such an integral part of the construction global datasets, such as for character data matrices, ZooBank, specimen databases.

Acknowledgment

This work has been supported by the US National Science Foundation and the German Deutsche Forschungsgemeinschaft.

REFERENCES

BHL. Biodiversity Heritage Library. Available on: <http://www.biodiversitylibrary.org>. 2007.

Brown Junior, W.L. Rhoptromyrmex melleus, brief characterization of. *Pilot Register of Zoolo*gy, v.13, p.1, 200?.

EOL. Encyclpopedia of Life. http://eol.org, 2007.

FORMICIDAE. Available on: <http://formicidae.myspecies.info>. 2007.

KNAPP, S.; POLASZEK, A.; WATSON, M. Spreading the word. *Nature*, v.446, p.261-262, 2007.

POLASZEK, A. A universal register for animal names. *Nature*, v.437, p.477, 2005.

SAUTTER, G.; AGOSTI, D.; BÖHM, K. A Combining approach to find all taxon names (FAT) in Legacy Biosystematics Literature Artikel. *Biodiversity Informatics* v.3, p.41-53, 2006.

WILBANKS, J. The Neurocommons. Available on: <http://neurocommons.org>. 2007.