# Playful Validation of Automatically Extracted Data

Francis Dierick, Philipp Dopichaj, Uwe Fleischer,
Andreas Heß, Andre Skusa, and Christian Maaß
Lycos Europe GmbH, Gütersloh, Germany
{francis.dierick,philipp.dopichaj,uwe.fleischer,andreas.hess,
andre.skusa,christian.maass}@lycos-europe.com

**Abstract:** We describe a community-based web site about German history, covering user interface and information extraction. The main focus, however, lies on creating and employing game approaches to validate and enhance the extracted data.

## 1 Introduction

The Semantic Web aims at annotating the World Wide Web to enable computers to understand the contents [BLHL01]. Ontologies are widely accepted in academia as a method to formally describe this additional knowledge. However, while *tagging* has taken off on the web, ontology editing has gone virtually unnoticed by common users, mostly because it is hard to do even for a specialist.

In the framework of the Alexandria project we are building a community website about famous people in German history based on Semantic Web technologies.[1] For this project, it is important to let the users adapt the ontology to suit their needs, but an important constraint is that the website should still be accessible to non-specialists. To achieve this, we developed new approaches that try to avoid direct contact between the users and the ontology. Our goal is to combine automated extraction methods with user knowledge in order to get most of the work done by algorithms and let the users validate the provided information in a convenient way. One of the hurdles in building a community site is providing a sufficient amount of interesting content for the community process to take off. This data should be extracted automatically, which entails a certain degree of inaccuracy that needs to be handled by user intervention in form of interaction.

We envision the interaction of the user with the ontology to take place within small casual games similar to those presented at the 'Games with a purpose' site[2]. We will show how automated extraction combined with games based on the principles of agreement, prediction and competition can help create the initial dataset for a large community website.

[1] Please register at http://www.alexandria-projekt.de to receive a private beta invitation.
[2] http://www.gwap.com


Figure 1: Screenshot of the Alexandria prototype showing three approaches to integrate extracted relations and facts into a page about a famous historical person. (1) relationship browser, (2) infobox, (3) human-improved fact based on computer-generated triple (highlighted).

## 2 Application

Our history website contains six main types of information:

1. Historical persons (e. g., Otto von Bismarck)
2. User-contributed articles
3. User-contributed images
4. Extracted structured data (e. g., date of birth)
5. Extracted relations (e. g., 'X is married to Y')
6. User-contributed 'facts' : short factual paragraphs about historical persons based on extracted triples.

All of this information is tied together by principles commonly found in community sites (e. g., tagging and voting). Our goal is not to create a typical community workflow, but rather to conceive and describe a process of seeding such a site with a large set of automatically extracted data. More specifically, we will describe how this initial seeding process can be enhanced by using casual games to validate and enhance the extracted data. Figure 1 shows a typical page as built after extensive usability testing. The practical application of the extracted data is outlined below.

1. *Extracted relationships* are shown in a relationship browser that is used for navigation and exploration. For ease of use, the density of information is kept low, but a full-screen version is available, providing sophisticated filtering functions.

2. *Extracted structured facts and images* are shown in an info-box which summarizes basic knowledge about a historical person.

3. The *five most important facts* are shown as short user-contributed paragraphs of text. At the core of each of these facts, there is a computer-generated triple (highlighted) which has evolved into the paragraph due to the site's unique collaborative process.

To populate the initial database, we extract *basic data* about the persons (e. g., dates of birth and death) from the German version of Wikipedia using an enhanced version of DBPedia [ABK+07]. We also extract *relationships between persons* by examining the link structure and the Wikipedia categories. This often yields a high number of relationships, making necessary ranking by importance.[3] The *type of the relationship*, e. g., 'colleague', 'married', is also of interest; it can be determined by looking at the direct context of the links or indirect context retrieved via web searches. We explored machine learning and pattern-based approaches on this context to qualify the type of relationship. To determine the *strength of the relationship*, we use search engine results and evaluate the number of hits for both names alone and together.

## 3 Games

Since all approaches for automatic extraction of data from Wikipedia have problems related to quality or performance, we combine the output of different approaches and then let users validate the information by using games.

Luis von Ahn describes the use of the human brain as processors as "human computation" and introduces the term "games with a purpose" [vA06].[4] The purpose of such games is to encourage users to collaborate on tasks computers are inherently bad at, e. g., image labelling or collecting common-sense facts. All these games have several things in common: they are *collaborative* and are thus based on *agreement* between multiple players. The *information objects are not changed* in the process; rather metadata is added to them.

### 3.1 Validation Games

The first step towards a human-readable representation of extracted information is to ascertain validity without changing the information object. A simple approach for validation is simply showing the information to the users and letting them decide if the given information is true or false[5].

We have applied such a simple feedback mechanism in a demonstrator of our relationship extraction technology called *Marry-o-Meter* [6]: a website that tries to automatically answer the question 'Are X and Y married?'. Along with the relationship answer (triple of the form subject – married – object) we present some basic information about the persons

[3] We explored the number of links or popularity on the web for ranking.
[4] Examples of such games are the ESP game [vAD04] and Peekaboom [vALB06]
[5] Such an approach is explored in the Cyc game: http://game.cyc.com
[6] http://www.andreas-hess.info/projects/marryometer/index.html

extracted from Wikipedia. We found that new users like to pick example queries from a list of suggestions like this:

- Brad Pitt ↔ Angelina Jolie
- Albert Einstein ↔ Mileva Maric
- Angela Merkel ↔ Joachim Sauer

We exploited this user behaviour by populating the list with sample queries that are of interest to our history website. We then provide feedback options to allow users to agree or disagree with the predicted relationship. We are developing more casual games based on this *principle of agreement* to help us validate basic facts about persons. 'Caught in history', for example, is a clue-driven game where users collaborate to find a certain historical person. By restricting the clues to template-based triples (e. g., X – died in the year – . . . ), the players unwittingly validate predicted triples. An example of gameplay could be:

Player 1 is caught in history by Napoleon Bonaparte and gives structured hints to Player 2, e. g.:

- Player 1: X was born in the year 1769 (triple: X ↔ born in the year ↔ 1769)
- Player 1: X lived in Elba (triple: X ↔ lived in ↔ Elba)
- Player 2: I think X is Napoleon Bonaparte

We found *template-based games* driven by the principle of agreement to do well for validating predicted triples without modifying the information objects themselves. These types of games are sufficient to seed the relationship browser and info-box in our history site (see Figure 1: 1 and 2).

### 3.2 Improvement Games

In some cases the extracted information objects need to be changed before they are presented to the public. We will explore two examples: *thumbnail image creation* and *improving human readability of triples*. The interaction with anonymous, unpredictable, human users plays a central role in making games with a purpose fun to play. By leveraging the efficiencies of competitive prediction markets we hope to evolve the crude triples into more human-readable units of knowledge.

For usability reasons it is desirable to restrict automatically extracted thumbnail images to pre-defined aspect ratios (e. g., the Wikipedia images in the info-box, Figure 1: 2). Automatically creating thumbnail images is hard and user input is typically required in order to create well-centred thumbnails. When asking multiple users to manually create thumbnails for a given image their selections tend to aggregate around a limited set of coordinates that can be easily grouped using a clustering algorithm.

We are developing a simple game called 'Thumbinator' where users are asked to create thumbnails with restricted aspect ratios. We exploit human curiosity by applying a scoring principle based on prediction: we give some information (e. g., an image) and a simple

Bismarck marry Johanna von Puttkamer

→ Otto von Bismarck married Johanna von Puttkamer.

→ Otto von Bismarck married the noblewoman Johanna von Puttkamer in 1847.

→ ...

Figure 2: Evolution of a simple computer-extracted fact.

question (e. g., 'create a square thumbnail' or 'is there a person in this picture?') and let users check how close their answers are to those of others. Game scoring is based on clustering of the responses. Anecdotal evidence shows that out of a corpus of 11000 Wikipedia images about 60 percent needs to be discarded for various reasons (inferior image quality, wrong image subject, ...).

*Triple evolution* aims at improving readability of extracted information. A given text can be represented as a sequence of triples conveying meaning. These triples can be automatically mined from existing texts; Powerset[7] uses this approach to summarize Wikipedia entries.[8] This extraction process produces triples that are sometimes hard to understand for humans.We use these crude triples at the core of a game applying prediction market principles. The player's goal is to take a triple, improve it and watch it grow in a *competitive* environment where only the best triples survive. An example of this triple evolution is given in Figure 2.

These evolved triples are then incorporated in the website as 'facts' (Figure 1: 3) in both human-readable and computer-readable form.

## 4 Conclusions and future work

We have looked at a community application that uses extracted information and relations about famous people in German history. Our work spans the range from actually extracting that data to verifying and displaying it. Extraction is based on Wikipedia content and explores approaches that can be used to determine how important a relation is for that person as well as techniques for determining the type of relationship. We have explored the principles of agreement, prediction and competition in casual games to validate extracted data and seed a community website with valid and interesting facts. We are currently evaluating the Thumbinator game on a corpus of 11000 images extracted from Wikipedia.

Although most of the work presented here has already been implemented in a prototype, a lot of refinement has to be done in order to achieve production-level quality and performance. In future work, we want to explore how a combination of these evolved triples can be used to create new human-readable knowledge.

[7] Powerset: http://www.powerset.com

[8] E. g., in Powerset, the sentence 'Bismarck married the noblewoman Johanna von Puttkamer (Viartlum, April 11, 1824 – Varzin, November 27, 1894) at Alt-Kolziglow on July 28, 1847.' is summarized by the 'triples' (they call them 'factz'): *Bismarck married Johanna von Puttkamer* and *Bismarck married noblewoman*.

**Acknowledgements.** The research presented in this paper was partially funded by the German Federal Ministry of Economy and Technology (BMWi) under grant number 01MQ07008. The authors are solely responsible for the contents of this work. We thank our colleagues at Lycos Europe who gave valuable feedback, especially Elica Savova for usability research and Jörn Schreiber for design work.

## References

[ABK+07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proc. 6th International Semantic Web Conference (ISWC 2007)*, pages 722–735. Springer, 2007.

[BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–43, 2001.

[vA06] Luis von Ahn. Games with a Purpose. *IEEE Computer*, pages 96–98, 2006.

[vAD04] Luis von Ahn and Laura Dabbish. Labeling Images with a Computer Game. In *Proc. CHI 2004*, 2004.

[vALB06] Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In *CHI '06: Proc. of the SIGCHI conf. on Human Factors in computing systems*, 2006.