Texty, texts at a glance

Jaume Nualart
Ludwig Boltzmann Institute
Kollegiumgasse, 2
A-4010 Linz, Austria
jaume@nualart.cat

ABSTRACT

Normally in visualization tools we find representations of the metadata of some documents, for example, pictures represented in a geographical map. Or authors producing texts in a timeline representation. There are also a lot of examples of pure text representation, based on the frequency of words or on the most frequently used words in a collection of texts. The idea behind Texty is to maintain the structure of the text and say something about the structure of knowledge in the text itself. With texty you can see, at a glance, in which area the text becomes more theoretical or more historical, for example.

Keywords

visualization, texts, knowledge, density, icons

1. INTRODUCTION

Visualization is hype and a lot of fashionable images and animations are appearing everyday. I recommend following in particular the reference page infosthetics [1]. Most visualizations are becoming more and more beautiful and attractive, art and design is here. But what I'm looking for when I create new visualization tools is the use of them. I call these beautiful visualizations constellations. When we look at the sky in a dark place at night we see beautiful images full of stars, planets, the moon,... But usually you need some knowledge to be able to understand what is there.

Taking this critical angle, we started the texty project thinking along the lines of tools to work with texts, mainly with collections of related texts: articles, documents reviews, etc. A tool to browse, read, research and understand archives of texts.

I wanted to create a text assistant as an interface for transversal reading and note-sharing with other readers. The text assistant is a huge project and is not yet completed. It will require more time, resources and people working on it. As a first step for this text assistant I started to imagine how we can represent a text as a small image, like an icon that represents some aspects of the texts. Here is when the texty idea appeared. The main idea of this small representation is to keep the structure of the text and say something about the structure of knowledge within the text itself. With texty you can see, at a glance, in which region the text becomes more theoretical, or more historical, for example.

In this article we explain how we implemented this process as researchers of the Ludwig Boltzmann Institute, a media art research center in Linz, Austria, during the period August 2008 to September 2009.

Texty cannot dynamically generate images of texts, it is a first case study for the texts of the Ars Electronica Jury Statements. The challenge for this next year is to create an application able to dynamically generate textys. I will inform readers about future works concerning texty in my blog nualart.cat

2. TEXTY: DEFINITION

Texty, texts at a glance, is a thumbnail image of an annotated text. It is a physical picture of a text in which you can see only parts of this text in the form of colored points. The colors refer to the categories of annotations in the text. For example, if you see red points this can mean that there are names of persons in the text. You can see where the references to persons appear in the text.

Texty does not tell you exactly what the text is about, but gives you a concrete idea about the structure of the text. You can read, for example, that a text is more theoretical in the initial paragraphs, or you can see that the text references works at the end of the text. You can see lists of items or even how many questions a text has.

Texty does not pretend to be a complex tool for understanding the semantics of a document or a collection of documents, but is a very light tool used to quickly gain knowledge about the contents.

The compromise is speed versus knowledge: what information can I obtain in a two-second viewing?

[1] http://infosthetics.com

Showing colored zones with meaning in a texty

Every colored point in a texty tells you that there is a word or an expression that belongs to a group of words, or specific vocabulary. The amount of points of a certain colour that one sees indicates the importance of this group of related words in the text. We can speak of a density of points, like points per square centimeter, or maybe a more precise and useful description: points per line, points per paragraph, points per text or more precisely, annotations per line. Obviously these are just density measures.

From text to texty

We started to use the expression Visual Knowledge Density (VKD) during the process of texty creation because we needed to have an idea about how many points we had in the texts. Probably it is not easy to find measures of VKD that can be used by different visualization and data management tools. So here we are just inviting other people in this field to start a discussion about this concept.

What we found interesting about the process is something related to the creation of the vocabularies representing knowledge fields. As ontologies, to create a thesaurus, apart from being a difficult task, it is also not a very unique job. Every author of an ontology will create a different one, and this is because knowledge is full of abstract concepts. Unlike other cases, for the creation of textys, it probably doesn't matter whether the vocabularies are the same or not since they cover a general list of words and expressions representing a specific study field.

3. TEXTY SHOWCASE: ARS ELECTRONICA JURY STATEMENTS

In our case, we wanted to generate textys for 91 texts we drew from the Ars Electronica Jury Statements. The texts are from 1987 to 2007 and cover several Golden Nica prize categories: u19 freestyle computing, Hybrid Art, Digital Musics, Interactive Art, Computer Animation, VFX Film, etc.

The text statements are explanations by the juries about their decision for each Golden Nica prize, the prizes of the festival Ars Electronica since 1987. The texts were selected because they have a similar structure and in partiular and more importantly, the texts belong to the same domain: media art.

The generation of the textys was preceded by a great deal of work in the creation of vocabularies and taxonomies for the media art field in order to annotate the texts. We used the work of G. Dirmoser [2], a specialist in semantics with more than 10 years of experience in the field. The way that Dirmoser creates the lists of words is not covered in this article. Following this, D. Offenhuber[3] annotated the texts using GATE software[4].

From Dirmoser's work we retrieved a list of words that are related to and are essential in media art, grouped in 5 categories: keywords, persons or institutions, artworks, awards and dates. We found 508 different annotations from 5.917 in total, that means that the words are repeated 10 times on average in the 91 texts.

Annotations per text: 65

Number of lines per texts (average): 200

Density of annotations: one annotation per 3 lines of text

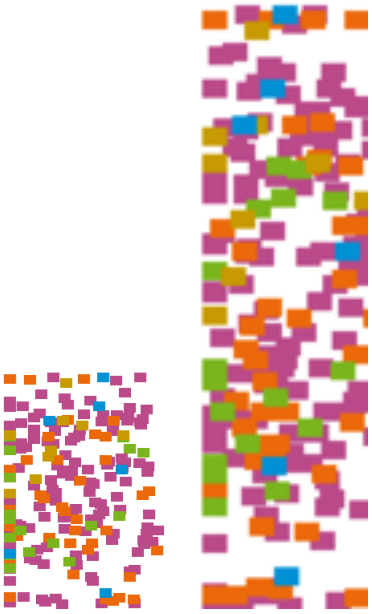
The textys are accessible here: http://vis.mediaartresearch.at/textass

4. TEXTY: THE SHORT TEXTY HOW-TO.

Texty, as a compromise between maximum knowledge and minimum time or effort, to be read it requires the legend for the selected texty view. The legend describes the type of annotation according to its color.

- Relative or absolute height: the shorter the icon the bigger points.

[2]SemaSpace - Semantic Networks as Memory Theatre http://residence.aec.at/didi/flweb/semaspace.pdf
[3]http://media.lbg.ac.at/en/content.php?iMenuID=87
[4]GATE home page: http://gate.ac.uk/



Relative Size Absolute Size

In the first tests, the relative size is more usable

The legend is necessary to understand the texty:

In this case the legend shows the strict name of each group of annotations. These names can be like:

Keyword	Theory	Essential words in media art field
Award	Prizes	References to festivals or prizes
Person	Social	References to persons and institutions
Date	Events	References to dates
Artwork	Works	References to works
default	-	In that case used for non clasified annotations

Some other vocabularies that have not yet been implemented could be:

History Texts dates, names, places,...

Chemistry Paper reference, keyword on a described field, element/formula, maths,...

Sociology Paper authors, references, keyword type 1, keyword type 2, keyword type 3,...

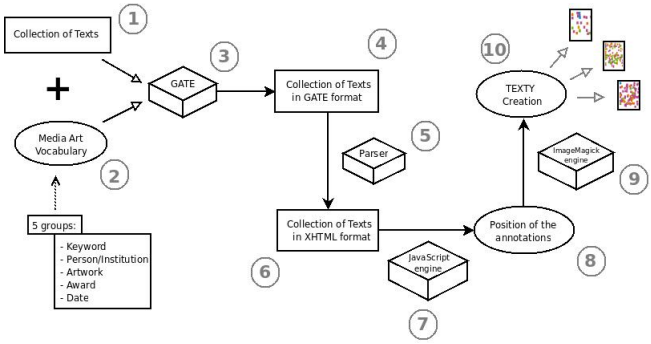
Personal Vocabularies: based on families of words created on the basis of the interests of the user.

Pure Text Structure1 question marks, list items, virgules,...

Pure Text Structure2 verbs, names, prepositions,...

5. TEXTY TECHNICALLY

5.1 Process



1. Collection of texts: we started with 91 articles in plain text format[5].
2. Media Art vocabulary: The vocabulary is a list of the most representative words in the Media Art field grouped in 5 groups: keywords, persons/institutions, artworks, awards and dates.
3. Annotator: we used GATE, General Architecture for Text Engineering[6]:
4. Collection of Texts in GATE format: we exported the annotated texts in XML-GATE format, where every annotation had a unique ID and similar and identical annotations had a group of related IDs.
5. Parser: The parser removed all the XML and added a simple XHTML format. Also we parsed all the grouped IDs and implemented a unique ID for all the synonyms. We also added a class name based on the kind of annotations according to the 5 groups mentioned previously.

[5]See point 3 texty showcase: Ars Electronica Jury Statements

[6]GATE http://gate.ac.uk - Infrastructure for Human Language Technology: General Architecture for Text Engineering or GATE is a Java software toolkit originally developed at the University of Sheffield beginning in 1995 and now used worldwide by a wide community of scientists, companies, teachers and students for all sorts of natural language processing tasks, including information extraction in many languages.
GATE comprises an architecture, a free open source API, framework and graphical development environment.
GATE community and research is involved in several European research projects including TAO and SEKT.

6. Collection of text in XHTML format: We imported the texts into a MySQL database and published them on a standard XHTML page.

7. JavaScript engine: we used JQuery[7] free software library with the plugin Dimensions[8] to be able to get the physical positions of every annotation of every text.

8. Positions of annotations: we get the list of the positions (x, y) of each annotations from a browser with default configuration[9] and this kind of annotation. We also obtained the height and width of the text itself. We did this using the standard client side JavaScript (Firefox 3.0). For the dynamic creation of texty, probably the best way would be to use server side JavaScript engines, such as the popular Rhino [10]

9. Image Builder: we used the ImageMagick[11] engine: Using the standard free software library ImageMagick we built several versions of the image (PNG, JPG, GIF). We also produced a second script using PHP to create the SVG versions of the textys.

10. Texty creation: we created several versions of textys and then we saved them in a database to be able to use them easily.

5.2 Annotations

Here, an annotated text means a text in which some words are marked using standard formats such as: RDFa, Embedded RDF or Microformats

In our first case we used a microformat as a simple XHTML in which an annotation looks as follows:

```
<span id="[uniqueID]" class="[KindOfAnnotation]">annotation</span>
```

The classes are mainly used to color the annotations. We will see a text like:

> "Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."

[7]Official website: http://jquery.com/ - Wikipedia definition: jQuery is a lightweight JavaScript library that emphasizes interaction between JavaScript and HTML. It was released in January 2006 at BarCamp NYC by John Resig.
jQuery is free, open source software Dual-licensed under the MIT License and the GNU General Public License.

[8]Website: http://plugins.jquery.com/project/dimensions - Extends jQuery to provide dimension-centric methods for getting widths, heights, offsets and more.

[9]Firefox 3.0 Browser under Ubuntu 8.10 GNU/Linux

[10]http://www.mozilla.org/rhino/

[11]Website: http://www.imagemagick.org/ - ImageMagick is an open source software suite for displaying, converting, and editing raster image files. It can read and write over 100 image file formats. ImageMagick is licensed under the ImageMagick License, a BSD-style license.

6. PLANS

Texty is ready to be applied to a second collection of texts and we will continue studying its utility working with texts: reading, browsing, comparing, filtering.

The main point we are interested in developing is the dynamic-texty. Adding new texts or re-editing texts of a collection and rebuilding the textys in server-side. We carried out very small experiments using the JavaScript server-side engine Rhino[12], to be able to get the real positions of the annotations. Anyway, there are more ways to get relatively good positions for the annotations using other techniques, such as counting the character positions in the text, to name the most simple one.

We have also been able to rebuild new texty templates, that is, using other vocabularies to generate the texty and give to the user several textys for each article.

One more feature for texty in the future will be a jumping-on-click to the text when clicking on the texty.

The dynamic texty will provide the path to integrate texty in the most popular CMS and Blog systems, where the number of entries and the complexity of origins of shared contents calls for tools in the field of visualization.

It also requires a development in the creation of personal vocabularies: personal collections of terms grouped in, we propose, a maximum of 7 groups (vocabularies). These vocabularies can be built automatically using the tags you use in your social bookmarks like delicious[13], or any other social network.

7. CONCLUSIONS

Texty, as you can see, is in the early stages of development. In this short paper we simply wanted to show you the potential of a very small tool and encourage you to use this idea in your projects.

Texty also wants to point out the importance of simplicity in the state-of-the-art information projects. Many projects are gathering complexity using more powerful computers instead of simple tools, shortcuts, hacks, silly ideas, jokes and dreams.

Keep it as simple as possible! This is not a call for primitivism, but a call for common sense.

[12]Rhino website: http://www.mozilla.org/rhino/

[13]http://delicious.com/