IST FP7 231507

D3.1 Report on Data Pre-processing v2.0

File name	PuppyIR-PuppyIR-D3.1-Report-on-Data-Pre-processing-v2.0.doc
Author(s)	TUD, EKZ, Museon, KUL, UT, UoS
Work package/task	WP3/ Tasks 3.2, 3.3
Document status	Final
Version	2.0
Contractual delivery date	30/09/2009
Resubmission date	30/09/2010
Confidentiality	Public
Keywords	data, content handling
Abstract	The report describes the data sets available through user parties EKZ and Museon for development and evaluation tasks, as well as the envisaged pre-processing approach. Also a series of external and/or public data sources with information suitable for children is reviewed. These could either be crawled from the internet or made available via search APIs. Version 2.0 is a resubmission reflecting the reviewers' recommendation to incorporate a more detailed section on implemented pre-processing measures.

Table of Contents

Executive Summary

The report describes the data sets[1] available through user parties EKZ and Museon for development and evaluation tasks, as well as the envisaged pre-processing approach. An inventory of potential useful data is presented together with the steps to be taken for the data to be suited for deployment in the PuppyIR research activities.

The report also reviews a series of external and/or public data sources with information suitable for children. These sources could either be crawled from the internet or made available via search APIs. These data sets may be used to test particular services and associated tasks, such as person findings, classification, summarization, etc.

The current overview is considered to be a dynamic document that will be regularly updated during the upcoming stages of the project. The data sets will serve as a source from which the main collection of documents and queries to be used in some of the other work packages will be drawn, and in particular will be input for WP5's deliverable D5.1 *Multimedia test collection*.

The report concludes with on overview how the identified data sources relate to the framework of variables expressed in D1.2 *Report on user requirements and scenarios*.

Version 2.0 reflects the reviewers' recommendation to incorporate a more detailed section on pre-processing measures implemented in collaboration between the user and technical partners (cf. subsection 4.3).

[1] The term 'data' is supposed to refer to both primary content and metadata. The latter includes both traditional metadata, user generated tags and hyperlinks.

Introduction

This deliverable surveys the main content/data sources available for research purposes of PuppyIR. It lists the pre-processing steps to be taken to turn the identified data sources into data suitable for the research activities to be carried out.

Two cases are distinguished: (i) internal data collections for which the IPR is owned by a consortium member and/or for which agreements with the IPR owners are likely to be feasible, and (ii) data available via external websites. For data sets of category (i) this report describes the steps to be taken in order to ensure that the data becomes available. For data sets of category (ii) the report describes the preprocessing steps (to be) taken in order to ensure that the data can actually be used.

The report is organised as follows: Section 2 summarizes the content and data sets available within the consortium through user parties AMC/Emma Kinderziekenhuis and Museon. Section 3 reviews public data sources with information suitable for children. The latter could either be crawled from the internet, or made directly available in PuppyIR research prototypes via search APIs. Section 4 describes the data handling and pre-processing steps. Section 5 relates the available data sources to the framework of variables identified in deliverable D1.2 *Report on user requirements and scenarios* (Annex I), and some concluding remarks are presented in Section 6.

Consortium Data Inventory

This section presents an overview of the content collections that will be provided directly by consortium partners (EKZ and Museon), or by organisations in the networks of the academic partners that are likely to agree to provide content under certain conditions.

1.1 Data from AMC/Emma Kinderziekenhuis

The Emma Kinderziekenhuis (Emma Children's Hospital) is a centre for paediatric health care. It offers a complete range of health care services for children from birth through 18 years of age, and is home to the paediatric medical training program in the Netherlands. The EKZ Patient Information Centre, Emmainfotheek (http://www.amc.nl/index.cfm?sid=300), provides the following resources:

- A library with books for children (and parents) about health, diseases, and hospitals. The library collection is summarised at Boeken voor kinderen (http://www.amc.nl/index.cfm?sid=1022), and includes a small collection of DVDs
- A collection of digital brochures for children, see Folders speciaal voor kinderen (http://www.amc.nl/index.cfm?pid=5192)
- A specific text with material meant to support school projects (called "Spreekbeurtboekje")
- Online information about a relatively new disease (new in The Netherlands), "sikkelcelziekte" (sickle cell disease), targeted at either children or teenagers

Of particular interest for the PuppyIR research tasks is the data archive consisting of information requests, maintained by the information centre. The 2008 collection, covering 260 information requests, is made available to the project; more years can be made available upon request. This resource categorizes the person making the request (patient or parent; age group) and the way the request was made (visit, email, phone call). EKZ has annotated most of the records with the information sources used to answer the questions. The 93 questions posed by patients have already been translated into English.

1.2 Data from Museon

Museon is a multi-disciplinary museum with collections in the field of natural history, history and archaeology, science and technology and ethnology. The information that is available reflects this heterogeneity. It originates from exhibition scenarios, catalogues and other publications or educational activities.

Part of this information is accessible online. A large part of it however is currently only available offline, but it will be made accessible online during the course of the project. After pre-processing of the content along the lines described in Section 4, the data will be brought online and merged with the materials that are already accessible online:

- www.museon.nl, the museum's own public website. Currently it contains about 100 articles of interest for PuppyIR.
- www.svcn.nl, a website with ethnology related information. Next to information about objects in the collections of the participating museums, it contains over 2,500 ethnology related articles.
- www.natuurinformatie.nl, a website that contains biology and geology related information. The Museon is one of the contributors to this website. The website contains over 6,400 articles.

- www.museumlessen.nl, a new website with educational materials based on museum collections.
- www.landvanhedenenverleden.nl, a Museon website with structured, interactive educational materials related to a limited number of subjects.

A very rough estimation is that the final dataset will contain about 750-1,000 articles

Besides this so-called "contextual" information, Museon of course will also make available a database containing information about the objects in its collection. This meta-information consists of text and images. Part of this collection information is accessible via the web (www.museon.nl and www.svcn.nl), but the largest part is not yet online.

The process of collecting, selecting and preparing the datasets has already started. The publication of the information is planned to take place in three different phases, reflecting the usage scenarios that have been defined in WP 1. The first phase will focus on the dinosaur theme, the second on the sustainability theme. They are expected to be finalized in the beginning of the second quarter of 2010. Publication of information that is more indirectly linked to the defined user scenarios will be ready around the middle of the year.

In the PuppyIR scenarios, also local PuppyIR services are foreseen. For these services it is expected that it will be necessary to prepare some additional datasets. This work will mainly take place in the second half of 2010.

For the purpose of the project, access logs of the Museon web server will be made available. With regards to external websites on which information from Museon is published and that can be considered as part of our dataset, we will investigate the possibilities to give access to the log files as well. With regards to museon.nl and svcn.nl it is already possible to use the Google tools to have information about the way in which the sites are used.

1.3 Data via academic partners

1.3.1 Textual sources

KU Leuven identified a corpus of Dutch texts about news events, written for children of different age groups.

The corpus consists of magazines Knuffel, Klap, and Kits, by publisher De Eenhoorn (Wielsbeke, Belgium; the Kits website is found at http://www.kits.be/). The magazines target age groups 7-8 (Knuffel), 8-10 (Klap) and 10-12 (Kits); the collection consists of volumes 1-18 of Knuffel and Klap and volumes 1-20 of Kits. The publisher will deliver the data in either low resolution PDF documents, or as quark files (QuarkXPress is a publishing application). From these files, we will extract the actual texts, to a standard format (plain text or XML).

KU Leuven is in the process of finalizing a license agreement, which can then be sent to the publisher. The agreement allows sublicensing the collection to the other partners of the consortium.

The articles are now grouped by the magazine they come from, which is equivalent to being grouped by the age of the target group. The next step is to align the articles from the different magazines: for each article that talks about a specific event, we align this with articles in the other magazines that talk about the same event. This way, we construct a parallel corpus: for each event, we have (in the best case) three different articles, for each age group.

Depending on the quality of the data, alignment of the articles on a sentence level could be considered. However, this requires a feasibility study on the data first. Alignment would in that case probably have to be done (at least partially) by hand.

It is possible to extend the resulting parallel corpus with other versions of the articles on the same events, by searching the Web. The articles found there will be written for adults. (IPR aspects require some attention when taking this approach.)

As for an English version, there are two interesting options. Reading A-Z.com (www.readinga-z.com), contains over 2200 books (or rather short stories), grouped into 26 levels. Weekly Reader (www.weeklyreader.com) publishes magazines much like De Eenhoorn does. They have several magazines for the different grades the students are in. However, upon inspection of the free samples on the website, it appears that the content in the different magazines is rather unrelated to each other. The September issue for grade 2 is about the constitution, the 3rd grade edition about space. The 'senior' edition (grades 4-6) has more content, but none of which is related to the other two magazines. Both companies that manage these websites are located in the United States.

Finally, another somewhat similar source is FirstNews, a British weekly newspaper for children (aged 7-14); see http://info.firstnews.co.uk/.

The decision to contact any of these options will be made later, depending on the research needs. For now we will focus on the Dutch dataset.

1.3.2 Video sources

While not finalized, the VideoCLEF evaluation initiative is planning to introduce some tasks that would match the PuppyIR objectives quite well.

In particular, one of the planned tasks concerns recommendation of videos by predicting their age level, appropriateness or inherent appeal using spoken, visual, and audio content as well as accompanying metadata. Ground-truth for this task is derived from parental guidance scores, user ratings or statistics on viewing frequency. Video data could be taken from http://www.kijkwijzer.nl/, but data availability for use within PuppyIR is still unclear. However, this VideoCLEF task is lead by parties that we have good working relations with: Stephan Raaijmakers (TNO), and Roeland Ordelman (Institute for Sound and Vision, aka Beeld & Geluid, and Universiteit Twente/HMI (*sic*)).

The second task concerns the identification of videos with high and low levels of dramatic tension. In particular, distinguishing between video content that causes the viewer to feel bored or entertained. Visual features, speech transcripts and metadata can all be used for this task. Human viewers will provide the ground-truth. The affect task is lead by Mohammad Soylemani (Geneva, Switzerland), who will specifically coordinate the relevance assessment phase.

TUD will monitor the developments, and notify the consortium where appropriate.

Public Data Sources

A large number of external data sources could be used to study content created for children, content generated by children, and content that can be used by children after certain pre-processing and filtering steps.

In the following subsections, we detail the main entry point for crawling the data by the lines starting with bold-faced "URL:", the language(s) of the content by the line starting with bold-faced “Language”, and the anticipated ease of collecting that data source by the line starting with bold-faced "Crawling:".

1.4 Knowledge-repositories for children

1.4.1 Simple Wikipedia and Simple Wiktionary

URL: http://simple.wikipedia.org, http://simple.wiktionary.org

There are 57,118 articles written by about 700 users on the Simple English Wikipedia and 11,707 articles written by around 43 users on the Simple English Wiktionary; where only a very limited vocabulary is used in writing the articles, like:

http://simple.wiktionary.org/wiki/Wiktionary:Extended_Basic_English_alphabetical_wordlist

(1500 words)

http://simple.wiktionary.org/wiki/Wiktionary:Basic_English_alphabetical_wordlist

(850 words)

The editorial process of these two sites does not guarantee that the content is suited for children, so a filtering stage is necessary should we decide to include this data in PuppyIR corpus creation.

Language: English

Crawling: Dumps of Wikipedia projects and their Simple versions can be downloaded for free from http://download.wikimedia.org

1.4.2 Wikipedia for Schools

URL: http://schools-wikipedia.org/

Wikipedia for schools is a selection of Wikipedia, targeted around the UK National Curriculum of about 5500 articles. The selection is vast, and covers core subjects but does not try to be uniformly detailed: for example, it has more depth on Llandudno, which is featured in the curriculum, than on other but similar places. Articles were chosen from a list ranked by importance and quality generated by project members. This list of articles was then manually sorted for relevance to children, and adult topics were removed. SOS Children volunteers then checked and tidied up the contents, first by selecting historical versions of articles free from vandalism and then by removing unsuitable sections.

Language: English

Crawling: free download from the project’s website

1.4.3 Wikijunior

URL: http://en.wikibooks.org/wiki/Wikijunior

The aim of this project is to produce age-appropriate non-fiction books for children from birth to age 12. These books are richly illustrated with photographs, diagrams, sketches, and original

drawings. Wikijunior books are produced by a worldwide community of writers, teachers, students, and young people all working together. There are around 900 articles.

Language: English

Crawling: Can be downloaded as a part of Wikibooks project from http://download.wikimedia.org

1.4.4 Wikikids

URL: http://www.wikikids.nl

A Dutch online interactive encyclopaedia for and by children containing 5,421 articles, 19 moderators, 184,000 users registered, including 132 schools.

Language: Dutch

Crawling: a proper download location is unknown, but the HTML pages can be crawled directly.

1.4.5 Vikidia

URL: http://fr.vikidia.org

Vikidia is a French interactive web encyclopaedia devoted to children both as readers and (not exclusively) editors, containing 6 602 articles written by around 1600 contributors.

Language: French

Crawling: location of dumps is unknown. It is possible to download HTML pages.

1.5 Content generated by children

1.5.1 Blogs/Microblogs

URL: http://www.twitter.com

Some microblogs are created by children, who can be identified via search over Twitter profiles (using TweepSearch). However, considerable number of blogs can be found only using ages starting from 10 in the phrase "I am …":

10: ~60, 11: ~130, 12: ~234, 13: ~483, 14: ~619, 15: 700

Language: English

Crawling: Using the Twitter Profile Search API:

http://dcortesi.com/2009/04/07/twitter-profile-search-api/

URL: http://www.wink.com

Metasearch over many social networks and blogging systems. Using the same search strategy ("I am 10, 11…"), we can discover many more identities at web-sites like Yahoo! 360, Myspace, LiveJournal, etc. Stats are:

10 years old: ~6,000, 11: ~9000, 12: ~16,000, 13: ~63,000.

Language: English

Crawling: WINK has no API and returns only the first 100 results. However, many additional search options are available; unfortunately, filtering for specific identities seems to require manual processing.

1.5.2 Stories

URL: http://www.kids-space.org/

Hundreds of stories written by kids from 4 to 15.

Language: English

Crawling: It is possible to crawl the data in standard manner.

1.6 Directories of web-sites for kids

1.6.1 DMOZ (Open Directory)

URL: http://www.dmoz.org

Largest human-edited directory on the web, containing around 5,000,000 web-sites in 590,000 categories. However, the English part (not under categories **Adult, Regional, World** and **International)** contains about 1,400,000 web-sites. There are around 31,000 English web-sites classified as appropriate for kids (and around 16,000 of such web-sites in other languages). Web-sites are classified by age: 3500 web-sites are only for kids of 12 and younger, 14,000 are also for 12-years-olds, 1000 are only for 13-15, and 27000 web-sites are appropriate for them as well, around 2500 of web-sites are only for teens of 16-18. Categories of Kids-specific and general directories often match and there is often an explicit link between them.

Language: English

Crawling: DMOZ directory can be downloaded in RDF format from the project's web-site. Web-sites should be crawled individually. Some of them may be found in the ClueWeb90 crawl of 50 billion pages distributed among participants of Text Retrieval Conference (TREC).

1.6.2 KidsClick!

URL: http://www.kidsclick.org

Around 6000 web-sites for kids categorized under hundreds of topics and also appropriateness for kids from different grades: 1-2, 3-6, 7+. Web-sites are as well classified as either containing a few, or many illustrations.

Language: English

Crawling: Can be downloaded by a simple crawler.

1.6.3 Other directories

http://www.cybersleuth-kids.com/

http://www.kidsites.com

http://www.kids.gov/

http://www.surfnetkids.com/

1.6.4 Bookmark-sharing web-sites

URL: http://www.delicious.com

People at bookmark sharing web-sites often tag sites appropriate for kids with the tag "kids" and other tags specifying their topics as well. Delicious reports about 475,000 bookmarks with this tag (and 253,000 for the tag "children"). It is however important to realize that using this data still requires to manually filter out web-sites containing content not appropriate for kids, as nothing is known about the reliability of the person who tagged the content.

Language: multiple languages

Crawling: Delicious API limits the number of web-sites per tag to 100. However, it is possible to use combinations of tags (kids + some tag + some tag) in order to retrieve more sites tagged for kids. Besides, there are several crawls of **delicious.com** on the web, which can be used to extract these web-sites. For example, the crawl which is shared by DAI-Labor group (See dai-labor.de) containing 130 millions bookmarks, made between 2003 and the end of 2007.

1.7 On-line Literature for Children

This section covers briefly different types of material that could fit into PuppyIR collection: children digital libraries, reference web site designed for children, reference web site for parents, librarians and educators; for each of these a representative example is described.

The first entry is a digital library for children, which follows the philosophy of the Open Library[2] project, an innovative and ambitious Open Access initiative by the internet archive team. Their aim is to make all published books, free from copyright, freely available online. All software they produce is freely access and available via their web site. The latter acts also as a library portal giving users access to all available titles plus an option to order a scanned version of those not yet enlisted. The International Children's Digital Library follows the same approach to book acquisition: all titles are copyright free, scanned, OCR-ed and proofread.

The second entry is a wide reference web site with links to a mixture of different types of resources: small collection of digital books, reference material, search-engines for children, games and fun activities.

The third entry is a reference web site maintained by a professional source with good width and breath in terms of information about children books and their authors.

1.7.1 International Children's Digital Library

The International Children's Digital Library (ICDL), http://en.childrenslibrary.org/, is a project initially created by an interdisciplinary research team at the University of Maryland (from College of Information Studies and the Human-Computer Interaction Lab) in cooperation with the Internet Archive. It aims at building a free digital library of children's books in original languages from around the world and supporting communities of children and adults in exploring and using this literature through innovative technology designed in collaboration with children for children. Books are searchable by a variety of ways specifically designed by and for children, such as location, age, genre, size of the book, novelty, keywords and language. Text and icons are used in a rich combination to help children during the searching and browsing phase. Results are shown as thumbnail icons with title and language. Once children have chosen the book they want to read, a brief summary plus extra information about it is presented on the screen; children can then read the book by clicking on each page image starting from the cover page. There is a mobile book viewer to read books on iPhones. Books are all free from copyright and could be added to the PuppyIR collection, after asking ICDL research team for permission, both as text and image version.

1.7.2 KidsSpace@The Internet Public Library

The Internet Public Library is a public service founded by the University of Michigan's School of Information, hosted by Drexel University's College of Information Science & Technology. Its children section at http://www.ipl.org/div/kidspace/ is a very rich reference web site with pointers to resources for children, educators and parents. Links are classified according to purpose, so

[2] at http://openlibrary.org/

there are links to reference resources such as encyclopaedias and dictionaries, one to the Reading Zone with links to repositories of children books, information about authors and books including curiosities and trivia, there are also sections on Health, Internet Search and Games. All are reference pages with links to mostly American web sites that have been checked and judged appropriate for young readers. There is also a Teen Space, http://www.ipl.org/div/teen/, for adolescents. This web site is managed, maintained and monitored by specialist librarians so it is a safe place for PuppyIR children to browse and search directly.

1.7.3 ACHUKA

ACHUKA, http://www.achuka.co.uk/index2.php, was established in 1997 and has been online continuously since then. It is a reference web site and has links to book titles, reviews, relevant blogs, and in general informative and opinionated material on children books. Its target audience are parents, educators and librarians. PuppyIR could crawl its content or simply make it directly available for browsing and searching.

1.8 Public Medical collections

Some of the PuppyIR usage scenarios involve medical information. There is no “one best place” for medical information. Several specialised websites offer information, with often a section for children. Crawling can however be difficult, as these websites tend to use flash and other highly interactive interfaces.

Additionally, EKZ lists websites with medical information for children, in Dutch, created by reliable organisations: EmmaInfotheek web directory (http://www.amc.nl/index.cfm?pid=4874). The list below indicates which websites are also included in the EKZ web directory.

1.8.1 Medikidz

URL: http://www.medikidz.com/

Medikidz is a flashy website with a super-hero theme. Children can make friends based on illness or medical procedure, take part in discussions, or write their own blog. Also offers a glossary of medical terms, from 'eye' to 'Villi'.

Language: English

Crawling: Some flash, most is standard

1.8.2 Bennopagina

URL: http://www.nvhp.nl/nvhp-bennopagina.htm

Short magazines for children on haemophilia (8 total - PDF format).

Language: Dutch

Crawling: The text would have to be extracted from the PDF files. No large amounts of information, but could be useful for evaluation purposes.

1.8.3 Astmafonds

URL: http://www.astmakids.nl/

Information and prevention/practical tips, discussion forum, stories of sufferers, jokes, games, ... No large amounts of information, but could be useful for evaluation purposes.

Language: Dutch

Crawling: Standard

This resource appears on the EKZ list of reliable websites for children.

1.8.4 VKS Kinderweb

URL: http://www.stofwissel.nl/kinderweb/index.php

Site about metabolism (diseases), both for children (PDF text in illustrated story format, games, ...) and adults. No large amounts of information, but could be useful for evaluation purposes.

Language: Dutch

Crawling: Flash and PDF, manual work

1.8.5 WKZ

URL: http://www.hetwkz.nl/

The children section contains information on the Wilhelmina Kinderziekenhuis (Wilhelmina children's hospital), diseases, narcosis, ... There is some material the children can use for assignments and presentations.

The teen's section contains a page with a dictionary of several rheumatism terms that could be used for explaining the terms to smaller children as well.

Language: Dutch

Crawling: Difficult. The pages are in flash, some downloadable documents are in Microsoft word format.

1.8.6 Edheads

URL: http://www.edheads.org/activities/knee/

Children get information from a surgeon on the procedure of knee surgery in an interactive comic. They have to answer questions and can perform their own surgery on the cartoon patient. Contains additional info, e.g. glossary.

Language: English

Crawling: Glossary is crawlable, interactive parts are not.

1.8.7 Habits of the heart

URL: http://www.smm.org/heart/heart/top.html

Information about heart and lungs in the form of animations, videos and classes with experiments.

Language: English

Crawling: Difficult

1.8.8 Starlight

URL: http://www.starlight.org/familiesshare/

This webpage contains stories, mostly written by parents and teens. The diseases vary.

Language: English

Crawling: Standard

1.8.9 Kankerspoken

URL: http://www.kankerspoken.nl/

This website collects citations by children as well as some information on the disease. There is also a forum where the children can talk about their feelings. The forum is divided in three parts, one for children under twelve, one for older children, and one for adults.

Language: Dutch

Crawling: Standard

This resource appears on the EKZ list of reliable websites for children.

1.9 Public Data Collections via OpenSearch

OpenSearch (http://www.opensearch.org) is a collection of simple formats for the sharing of search results. The OpenSearch description format can be used to describe a search engine so that it can be used by search client applications. OpenSearch response elements can be used to extend existing syndication formats, such as RSS and Atom, with the extra metadata needed to return search results.

OpenSearch provides several flavours of interfaces:

1. the before-mentioned RSS and Atom feed like interface, leading to rather traditional search results with a title, some description and a url;
2. search suggestions, which provides only suggested queries. This is for instance used in Wikipedia when you type the first characters they're automatically expanded in a drop down box;
3. visual search suggestions, much like the RSS feeds, but meant to be suggestions as in Wikipedia, with additional things like images (hence, visual search).

There is support for this technology from the major players, and the support is growing quickly. Current versions of the major web browsers, such as Firefox 3.5 and Internet Explorer 8 operate as clients for OpenSearch, and major search engines and sites, for instance Yahoo!, Twitter, Amazon, and Wikipedia, support at least some flavour of OpenSearch search results to act as search servers.

PuppyIR will provide components that act as clients for OpenSearch services. Some of the public data collections in English or Dutch, targeted at children that contain trusted, moderated information for children (without commercial incentives), also support OpenSearch. Unfortunately, at the time of writing this deliverable, we know only few sites. Examples include:

- http://en.wikibooks.org/wiki/Wikijunior: a part of the English Wikibooks wiki that targets children; see 2.1.3 above;
- http://www.spelenboek.nl: a collection of recreational games in Dutch for children and families.
- http://wikikids.wiki.kennisnet.nl: a Dutch encyclopaedia for children of age 8 to 13.

These three sites support OpenSearch because they use Media Wiki, the wiki implementation that is developed for, and used by, Wikipedia. There are hundreds of thousands of web sites that run Media Wiki. For instance, wikia.com (http://www.wikia.com), contains over 50,000 community-created wikis, many of which target children (There are at least 7 English Wikis sites on Dinosaurs: http://dinosaurs.wikia.com, http://dino.wikia.com, http://dinosaurus.wikia.com, http://dinosaurika.wikia.com, http://dinosaurspedia.wikia.com, http://dinosaurfactsfile.wikia.com, http://trex.wikia.com). Many other sites use Media Wiki as well (amongst others SourceForge.net). A list of some of these can be found on the Media Wiki site (http://www.mediawiki.org/wiki/Sites_using_MediaWiki), but this is only the tip of the iceberg.

1.10 Involving Children in Building PuppyIR collections

Children are increasingly playing a central role in the design and evaluation of interfaces and systems for them to use (Guha et al, 2004; Markopoulos et al, 2008; Mazzone et al, 2008). Thus, we believe that they should also be directly involved in the construction of PuppyIR collections.

We propose to run an informal study where children 8 -12 will be asked about their favourite web sites and books. In order to run this survey as quickly and smoothly as possible, it will be conducted among children we have direct access to, that is, relatives and their friends, so that parental consent is implicit and there will be no need of disclosure for researchers involved (only required for UK studies, as in D1.1). A light procedure for the internal approval of the study will be adopted, as requirements on children will be kept to a minimum.

Children will be invited to contribute by providing addresses of favourite sites together with a brief description and possibly motivations behind preference. It has been observed that native digital children are naturally inclined to share this type of information informally among friends and class mates. PuppyIR will encourage children to submit their contribution to a public bookmark sharing site like del.icio.us, with common PuppyIR-child tag.

Equally informal conversations/interviews will be run on site with participants in order to get extra information on motivations and trends and enquire about the use and preferences regarding non-web based material such as books, encyclopaedia, etc. As the partners in the consortium come from different EU countries, it is expected that this exercise will provide us with a variety of input to paint a picture across languages and countries.

Data Handling

In order to streamline the access and use within the consortium of the data and/or content sets described above, a handling procedure has been designed for the two categories of materials that will be input for the tasks to be carried out: raw data and pre-processed data.

The Universiteit Twente (UT) has created a repository through which selected data will be available to all consortium partners.

1.11 Collection procedure

The following steps are distinguished for making information available for the consortium:

1. *Collection of materials*:
   a. internal data:
   Resources available via consortium members or directly provided to a partner in a variety of forms and formats (texts, image, audio, video) are collected.
   b. external data:
   content at external sites crawled or accessed via search API
2. *Selecting materials*
   Not all identified information will be useful for the project. Part of it may not be suited for the age groups targeted by the project, part of it may be outdated.
3. *Preparing the datasets*
   Since the materials were most likely developed for a specific aim (e.g., in the museum case this could be exhibition and education), it is expected that some additional processing or updating may be required for optimal support of on or more of the chosen use scenarios (cf. D1.2).
4. *Distributing materials*
   The materials are uploaded to the UT repository and available for use within the consortium.

1.12 Repository Management

Data is initially stored in their raw native format, accompanied with a “README” file for specification of the metadata.

The metadata README file will specify at least the following:

1. Metadata formats (XML schema, DTD, Dublin Core, etc.)
2. Details on the origin of the data (IPR/copyright holders, creation date, crawl date and script used, url),
3. Details on who can use the data, for what purposes and under which conditions (research purposes only or wider use, consortium members only or wider public also, signature of IPR license needed or not, etc.)
4. Information on the data format (text, video, audio), encoding, and language(s)

Individual project partners involved in research tasks, software development tasks, or the development of demonstrators are responsible for any required processing, conversion, and enhancement (e.g. filtering) of the data. Data that has been pre-processed and/or enhanced will be (re)submitted to the UT repository. XML is the most likely format for pre-processed and enhanced versions of the data sets. Pre-processed/enhanced data is accompanied by the same type of README file as the raw data, but in addition it specifies:

5. Pre-processing info: (pointers to) scripts for conversion scripts, tokenization, filtering, named entity extractors, etc.).

1.13 Data pre-processing

Different types of data pre-processing are needed for different applications. This document will detail only the shared data pre-processing applicable for all usages of data in the consortium. Work-package or research-question specific types of pre-processing will be addressed in other documents (e.g., the deliverable on information extraction techniques (D3.3) is likely to specify additional pre-processing steps to support the algorithms developed in WP3).

The basic principle in all pre-processing is the conversion of original data sources to an XML format – preferably one that is easily mapped to an RSS feed for access through OpenSearch API. XML data can always be indexed by the open-source XML IR system PF/Tijah (partially developed by consortium partner UT), a full-text XML retrieval extension of MonetDB/XQuery. PF/Tijah supports XQuery extended with ranking on text nodes, and provides all standard IR pre-processing tools out-of-the-box, e.g., stemming and stopping; also, the transformation of the source XML to a different desired output format (like RSS) is straightforward using XQuery.

This take on data pre-processing is best illustrated using the EKZ collection, where the data provided by the EKZ has already been pre-processed at the UT. First, the internal EKZ database (in Microsoft Access) has been exported as tab-separated file. A custom Perl script is used to convert this database export into a canonical XML format that is easily transformed into an RSS result stream. For better result display, the collection has been augmented with images from the internet for book covers, dvd covers, etc.

A demo of an OpenSearch compliant search engine of the pre-processing result (using PF/Tijah to manage the resulting XML) is available at http://pathfinder.cs.utwente.nl/puppyir/.

The nature of the data of the Museon demonstrator are quite diverse, consisting of text, images, video and sound clips, both with and without metadata. Part of it is already available on websites, part is unpublished and part needs pre-processing. Here, we detail the data handling and pre-processing that has been planned by Museon for the two demonstrators.

The *interactive quest* starts at a multi-touch table. The users of the application begin by determining the subjects the quest should be about, by selecting images that represent exhibits. This step requires (manual) mapping of images to subjects. The quest itself is composed by retrieving the relevant questions from a MySQL database, where questions are already mapped to exhibits. These questions are based on information that is already available in the museum (and will be used for the second demonstrator as well). During the quest, players collect objects that they bring back to the multi-touch table for an end game. This end game consists of mapping these objects with related concepts/keywords by the players. To enable this mapping some data pre-processing is needed: objects are tagged with keywords (in an Excel sheet). The *multi-touch table demonstrator* is also related to the museum's permanent exhibition, enabling users to search through information related to the exhibition's subjects. For this application, the text, image, sound and video data that have been produced for exhibition purposes are made available. Text materials are made accessible by publishing them on a website. Image materials will be connected to the textual information on the one hand, while on the other hand they will be enriched with information from the museum's collection database. The raw data will be delivered in XLS-format. Video and sound will be annotated; the raw metadata will also be delivered in XLS-format.

From the project's perspective, the resulting data are considered the raw Museon data to be uploaded to the repository discussed in sub-section 4.2. Similar to the EKZ case, the collection of web-pages, the relational database with questions and the accompanying Excel spreadsheets are combined in a unified XML format (e.g., using a few Perl scripts).

Relationship to D1.2 (User requirements report)

Deliverable D1.2 *Report on user requirements and scenarios*, presents a framework of variables that play a role in the user scenarios. For some of them we have performed a preliminary assessment of the degree in which the data and content sets identified can be considered balanced.

With respect to *user characteristics*, many sources have been identified for different age groups. Gender differences are less clear in the current data sets: which of the sources are more suited for boys and which for girls we do not know yet.

With respect to *knowledge and language skills*, it has been concluded that sufficiently many resources have been identified.

The combination of the EKZ requests archive (annotated question-answer resource pairs) with the Museon web access logs (and therefore http-referrer information, containing the web search engine queries that pointed the user to these pages (see [Antonellis *et al.*, 2009]) should give an indication about the *search skills*.

For *evaluation skills / critical thinking* skills, the authority, reliability and recency variables are partially covered by the EKZ and Museon web directories. Especially for Museon, data pre-processing is however not completed yet. Given the current availability of content however, we foresee no problems for the creation of appropriate demonstrators.

A more or less open issue is how to acquire reliable information about children as content creators. Tagging data generated by children are needed for the development and demonstration of PyuppyIR information services in general, and for the integration of associated tasks like expert and person finding in particular. Based on the current overview clearly some targeted effort is needed for the collection of social tagging data. Section 3.2 provides a good starting point. Some partners have already made use of some of data described in that section.

Concluding remarks

The report identified a wide variety of data sources that can be used in the project. In addition to data sets owned by the user parties, many useful public sources have been identified, including a good amount of collection and/or sites that were designed especially for children. The project could crawl these sources to obtain snapshots, or, for a subset of them, to access the data through the relatively new OpenSearch API.

The current overview is considered to be a dynamic resource that will be regularly updated during the upcoming stages of the project. Based on an assessment of the overview it was concluded that the project should undertake additional efforts to collect social data (content tagged by children).

The current and future data sets will serve as the main collection of documents and queries to be used in some of the other work packages, and in particular will be input for WP5's deliverable D5.1 *Multimedia test collection*.

The PuppyIR consortium will consider to create a bookmark-sharing community for promotion of kids-sites, or to build the tools that would help existing communities establish such functionality, through better classification/description of their resources.

References

Antonellis, Ioannis and Garcia-Molina, Hector and Karim, Jawed (2009) *Tagging with Queries: How and Why?* In: Second ACM International Conference on Web Search and Data Mining WSDM 2009, Late Breaking Results Session, February 9-13, 2009, Barcelona, Spain.

Guha, M. L., Druin, A., Chipman, G., Fails, J., Simms, S., & Farber, A. (2004). Mixing Ideas: A new technique for working with young children as design partners. In Proceedings of Interaction Design and Children (IDC'2004). College Park, MD, pp. 35-42.

Markopoulos, P., Read, J. C., MacFarlane, S. J, and J Hoysniemi, (2008) Evaluating Interactive Products with and for Children, San Francisco: Morgan Kaufmann Publishers.

Mazzone, E., Read, J.C. & Beale, R. (2008), Understanding Children's Contributions during Informant Design. The 22nd BCS British-HCI 2008, Liverpool, UK.