



WeKnowIt

Emerging, Collective Intelligence for Personal,
Organisational and Social Use

FP7-215453

D6.5.2

Final data sets collection and integration

Dissemination level:	Confidential
Contractual date of delivery:	Month 30, 2010-09-30
Actual date of delivery:	Month 31 2010-10-15
Workpackage:	WP6 Architecture and Integration
Task:	T6.5 Data and content collection
Type:	Report
Approval Status:	Approved
Version:	02
Number of pages:	26
Filename:	D6'5'2.tex

Abstract

This deliverable deals with the final collection, exchange and repackaging of multimedia content (text, images, video, and speech), user and user interaction data to be used within both use case scenarios as well as for the evaluation of research performed within the scope of the project. Content sources enable semantic content analysis and metadata extraction, while user data enables social network analysis and trend detection facilitating the extraction and generation of Collective Intelligence.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.



co-funded by the European Union

History

Version	Date	Reason	Revised by
01	2010-10-06	First draft: collection of input from various partners	Börkur
02	2010-10-14	Finalized front matter and addressed comments from internal review to make ready for submission	Börkur

Author list

Organization	Name	Contact Information
Yahoo!	Börkur Sigurbjörnsson	Phone: +34 93 183 8823 Fax: +34 93 183 8901 E-mail: borkur@yahoo-inc.com
All Partners	All partners contributed to this deliverable by describing the collections they have used in the project	Phone: N/A Fax: N/A E-mail: N/A

Executive Summary

In this deliverable we discuss the final datasets used in the WeKnowIt project. Due to the heterogeneous nature of the tasks addressed within the project the partners collected custom fit data that helped them solve the task at hand. The different datasets do, however, come together through the integration of the corresponding services in the prototypes for the two different use cases.

We divide the datasets into three broad groups:

1. Datasets used in the Consumer Social Group Use Case (Chapter 1),
2. Datasets used in the Emergency Response Use Case (Chapter 2),
3. Datasets used for evaluating research activities within the projects (Chapter 3).

This should only be considered a rough categorization of the datasets since in many cases the datasets used for creating services for either of the use cases were also used to evaluate the research underlying the services. Similarly, some of the research evaluation datasets were used to improve techniques underlying the services used in the prototypes, although the datasets themselves are not used directly in the prototypes.

The datasets for each of the broad groups span a range of different data types. Table 1 gives an overview of the data types used. The section numbers refer to the sections where the corresponding dataset is described.

	CSG	ER	Research
Text/Tags	1.1 1.2 1.3 1.4 1.5 1.8	2.1 2.2 2.4	3.1 3.2 3.3 3.5 3.6 3.9 3.10 3.11
Images	1.2 1.3 1.7 1.8	2.1 2.2	3.2 3.4
Audio/Video		2.1 2.2	
Communications	1.6	2.1	3.8
Communities			3.7
Entities/Gazetteers	1.1, 1.2	2.3	3.9

Table 1: Overview of data types used in the different use cases and research evaluations.

The goal of this deliverable is not to describe the datasets in great detail but to give an overview of the datasets used and pointers to deliverables and publications where more details can be found about the datasets and their usage.

Abbreviations and Acronyms

ABC	Australian Broadcasting Corporation
API	Application Programming Interface
BBC	British Broadcasting Company
CCTV	Closed Circuit Television
CSG	Consumer Social Group
ER	Emergency Response
Exif	Exchangeable image file format
MPEG	Moving Picture Experts Group
POI	Point of Interest
Q&A	Questions and Answers
REST	REpresentational State Transfer
SCC	Sheffield City Council
SVN	Subversion
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
WebDAV	Web-based Distributed Authoring and Versioning

Table of Contents

1	Consumer Social Group Use Case.....	8
1.1	Wikipedia Points of Interest	8
1.2	ViRAL photos	9
1.3	Hybrid image clustering.....	10
1.4	Flickr Events	10
1.5	Flickr tag recommendations	11
1.6	CSG Delicious Favorites	11
1.7	CSG Flickr Uploads	12
1.8	Grand Travel Challenge.....	12
2	Emergency Response Use Cse	13
2.1	Sheffield City Council Emergency Planning data	13
2.2	ABC News articles	13
2.3	Yahoo Geoplanet/OpenStreetMap Mapping	14
2.4	Flickr ER Sheffield.....	15
3	Research Evaluation Datasets	16
3.1	FLICKR-1M	16
3.2	Flickr Datasets with visual features	17
3.3	Flickr dataset with temporal info	17
3.4	Flickr Logo database	18
3.5	Flickr Data for Sheffield/Cardiff/Cambridge areas prior to 2010	18
3.6	Sheffield City Council major incidents 2006-9	19
3.7	Twitter community dataset	19
3.8	Communications structures.....	20
3.9	Wiki/DBPedia	20
3.10	BIBSONOMY-200K	21
3.11	DELICIOUS-7M	21
4	Data exchange.....	23
5	Conclusions	24
6	References.....	25

List of Figures

1	Example text article from the ABC News dataset	14
---	--	----

List of Tables

1	Overview of data types used in the different use cases and research evaluations.	4
2	Datasets used in the Consumer Social Group Use Case	8
3	Datasets used in the Emergency Response Use Case	13
4	Datasets used in the Research Evaluation	16

1 Consumer Social Group Use Case

In this section we will summarize the datasets used in services used by the WeKnowIt CSG prototypes. This contains various type of information, such as textual descriptions of Places and Points of Interest, photos and photo annotations. The data was collected using social media sites such as Wikipedia and Flickr. Various user-generated datasets have also been generated by the user interaction with the CSG prototypes. This includes marking places or points of interests as favorites; uploading photos to Flickr groups during WeKnowIt field experiments; uploading photos to Flickr directly through the mobile web application.

Table 2 lists datasets used in the Consumer Social Group Use Case. Each dataset will be described in successive sections.

Dataset	Provider	User
1.1 Wikipedia POIs	Yahoo!	CERTH-ITI, TID, Yahoo!, SMIND
1.2 Viral photos	CERTH-ITI	CERRH-ITI, Yahoo!, SMIND
1.3 Hybrid image clustering	CERTH-ITI	CERTH-ITI, Yahoo!
1.4 Flickr events	Yahoo!	UKob, Yahoo!
1.5 Flickr tag recommendation	CERTH-ITI	CERTH-ITI, Yahoo!
1.6 CSG Delicious Favorites	Yahoo!	TID, Yahoo!
1.7 CSG Flickr Uploads	TID	TID, Yahoo!
1.8 WeKnowIt Grand Travel Challenge	All partners	CERTH-ITI, Yahoo!

Table 2: Datasets used in the Consumer Social Group Use Case

1.1 Wikipedia Points of Interest

We process the full Wikipedia dump¹ and extract articles containing geographic coordinates. The articles are linked to an external geographical database – Yahoo! GeoPlanet² to obtain a mapping from the Wikipedia article to the surrounding places (towns, states and countries). This mapping can also be reversed and for a given place we get a list of POIs (Wikipedia articles) that fall within its boundary [3, 24]. For a given place we use our GridFaces entity ranking system to rank the POIs using query log data and Flickr photo annotations [24, 23]. The POI information and ranking is exploited in the WeKnowIt travel information services and contains information about 268,946 points of interest in 110,835 places – 107,165 towns, 3,440 states, and 230 countries.

Apart from the POI information, we also extracted information about article/section/subsection co-occurrence of places and landmarks within the whole Wikipedia corpus and this information was used in the POI recommendation service (See further: D3.3 Section 2).

TID uses Wikipedia categories to filter search results. Each Wikipedia article is tagged with additional metadata that describe its content for further searches. The recommendatory system has a subset of these tags to perform accurate searches and filter results. The information has been dumped from Wikipedia and related to the points of interest of Barcelona. In the upcoming modifications of the WeKnowIt Wikipedia POI services the category information will be expanded beyond Barcelona and included for all locations.

¹<http://download.wikipedia.org/>

²<http://developer.yahoo.com/geo/geoplanet/>

Deliverables: D3.3 Sections 2, 3 and 4.

Publications:

S. Diplaris, A. Flores, B. Sigurbjornsson, N. Tintarev, M. Escriche, R. van Zwol, Y. Kompatsiaris. Collective Intelligence in Mobile Consumer Social Applications. In 9th International Conference on Mobile Business (ICMB 2010) [3]

R. van Zwol, B. Sigurbjornsson, et al. Faceted exploration of image search results. WWW '10: Proceedings of the 19th international conference on World wide web. 2010 [24]

R. van Zwol, L. Garcia Pueyo, M. Muralidharan, B Sigurbjornsson. Machine learned ranking of entity facets. SIGIR '10. 2010 [23]

Services: Travel information services (WP3_GetPOIs, WP3_GetInfo, etc.), WP3_POIRecommender. The services are used both in the desktop and mobile travel prototypes as well as the WeKnowIt Image Recognizer³.

1.2 ViRAL photos

Several datasets were collected and used in ViRAL⁴.

a) 1,117,059 Flickr images along with their geo-tags, tags and titles. Downloaded using geographical bounding boxes around the city center of 23 European cities.

b) List of 3112 landmarks from Geonames with corresponding Wikipedia articles. Only landmarks in the area of the predefined bounding boxes of the 23 cities. Collected using Wikipedia Geocoding Webservice⁵.

c) List of 8186 geo-tagged Wikipedia articles which are also related with landmarks. Only landmarks in the area of the predefined bounding boxes of the 23 cities. Collected with the service described in WikiProjekt Georeferenzierung⁶.

A part of the images taken from Barcelona city are annotated in order to be used in the evaluation of the visual retrieval, location and landmark recognition. Annotated set consist of 1081 Flickr images taken in Barcelona. Images are divided into 35 groups depicting the same building/scene. 17 of the groups depict landmarks and 18 non-landmarks. Each group is also associated with the location of the building/scene on the map. The 17 landmark groups are also related with the corresponding Wikipedia article(s).

Deliverables: D2.1.2

Publications:

Y. Avrithis, Y. Kalantidis, G. Toliás, E. Spyrou. Retrieving Landmark and Non-Landmark Images from Community Photo Collections. In Proceedings of ACM Multimedia (MM 2010), 2010. [1]

³<http://www.weknowit.eu/wkiimagerecognizer/>

⁴<http://viral.image.ntua.gr/index.php>

⁵<http://www.geonames.org/export/wikipedia-webservice.html#wikipediaSearch>

⁶http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Wikipedia-World/en

Y. Kalantidis, G. Toliás, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, S. Kollias. VIRaL: Visual Image Retrieval and Localization. In *Multimedia Tools and Applications* (Submitted), 2010. [11]

Services: (a) is used in WP2_VisualAnalysis service. All (a),(b) and (c) are used in WP2_TagProcessing service. Both services are used in Fannr, the post travel photo annotation prototype. Both services are also used in the WeKnowIt Image Recognizer.

1.3 Hybrid image clustering

Dataset of 128,714 images along with their geo-tags, tags and titles collected from Flickr. All images are geo-tagged and were taken in Barcelona. In order to collect the images from Flickr a query requesting only geo-tagged images was used, constraining the search in a rectangle area around the city center. All images uploaded until 3/06/2009 were downloaded.

Deliverables: D3.3

Publications:

S. Papadopoulos, C. Zigkolis, G. Toliás, Y. Kalantidis, P. Mylonas, Y. Kompatsiaris, A. Vakali. Image Clustering through Community Detection on Hybrid Image Similarity Graphs. In *ICIP 2010* [19].

Services: To be used in the service WP3_WP2_Hybrid_ImageClustering. A preliminary version of the service is already integrated in the desktop travel prototype and is planned to be integrated into the mobile travel prototype.

1.4 Flickr Events

Flickr dataset used for event detection. Contains geocoded photos from Barcelona. The dataset contains 103146 geotagged images from the city of Barcelona (a spatial bounding box was applied to remove irrelevant content from other areas). The tag vocabulary consists of 27383 significant tags (after removal of too infrequent, misspelled, and stopword annotations). All resources are associated with timestamps. This allows for joint analysis of spatial knowledge and temporal distribution of content for the purpose of event detection. Although the overall density of the geotagged content in social media is still rather low, preliminary experiments with this medium-sized dataset have demonstrated the ability of Bayesian methods to extract event-like clusters of content based on spatial and temporal information.

Deliverables: N/A.

Publications:

S. Sizov and A. Ens. EventFolk - Automatic Event Detection in Social Media. In *Datenbank-Spektrum*. 2010. [20]

Services: The solution has not been implemented as a regular service for the WeKnowIt infrastructure, but the Barcelona instance has been integrated in the CSG pre-travel prototype. Although the overall usefulness of the proposed approach has been demonstrated, the insufficient density of the geospatial metadata in social media does not yet allow for efficient and stable

event detection in a general setting, beyond the borders of few major tourism attractions like New York, London, or Barcelona. The number of geographically annotated images is supposed to increase in future as more devices will be able to capture the spatial information.

1.5 Flickr tag recommendations

A set of 207,750 images collected by querying Flickr with a geo-query centered on Barcelona. These images were uploaded by a total of 7,768 users. We first processed the image tags in the following way: We filtered very long and very short tags, tags that consisted of both numeric characters and letters, as well as tags from a manually created black-list. We also merged tags that were lexically very similar to each other as expressed by the Levenshtein distance. This resulted in a total of 33,959 unique tags, and 173,825 images tagged with at least one of them. Subsequently, we formed the tag-image list index and removed tags used in more than 350 images. Examples of such tags are Barcelona, Spain, Catalunya that can be considered uninformative for the particular dataset. This reduced the unique tags to 33,367 and the images tagged with at least one of them to 120,742. Furthermore, out of the original set of 207,750 images, there were 195,308 geotagged images.

Deliverables: N/A.

Publications:

S. Papadopoulos, C. Zigkolis, S. Kapiris, Y. Kompatsiaris, A. Vakali. ClustTour: City Exploration by use of Hybrid Photo Clustering. In Technical Demonstration session of ACM Multimedia 2010. [18].

Services: Powers the tag recommendation service used in the Fannr the post travel photo annotation prototype. The service is also used in the ClustTour prototype, developed by CERTH-ITI.

1.6 CSG Delicious Favorites

In the Consumer Social Group pre-travel prototype, users can add places and points of interest as favorites. One important factor is that the users can choose to save their favorites using the Delicious⁷ bookmarking service. This means that their favorites are available to the rest of the WeKnowIt consortium (and the general public) and can be used for synchronizing favorites between the desktop and mobile applications. Furthermore, the resulting dataset can be used for various research evaluation tasks.

Deliverables: N/A.

Publications: N/A.

Services: The favorites are available through the Delicious API and are used to communicate between the desktop and mobile travel prototypes.

⁷<http://delicious.com/>

1.7 CSG Flickr Uploads

As part of the CSG mobile application, users can connect to its Flickr account to upload images of the POIs they are visiting. The uploaded images are published publicly with, at least, two tags associated. The first one is WeKnowIt and relates the multimedia element with the project. The other one is a number that corresponds with the identifier of the Point Of Interest (POI). In future iterations of the CSG case study application, the image will be processed to add auto-generated tags.

Deliverables: N/A

Publications: N/A

Services: The service to upload and search images on Flickr is included as part of the mobile application using the public Flickr API. In other iterations of developments, images will be automatically tagged with the Image Analysis service before uploading them to Flickr.

1.8 Grand Travel Challenge

A collection of photos taken by WeKnowIt project members during the WeKnowIt Grand Travel Challenge held in Barcelona in March 2010. The WeKnowIt members explored the city with the aid of the WeKnowIt mobile web application and took photos along the way. The users then uploaded the photos to Flickr, added them to a group dedicated to the event⁸ and annotated them with various metadata. All in all 647 photos were added to the group by 14 users and annotated with 1669 tags (not unique).

Deliverables: To be included in D2.3 Social Media Intelligence.

Publications: N/A.

Services: This data is implicitly available as a "service" through the Flickr API method for obtaining photos in a group. In addition, this dataset is going to be used as a test dataset in the WP2_GetContentBasedWeight service implementation.

⁸<http://www.flickr.com/groups/weknowitgc/>

2 Emergency Response Use Cse

The data used in the Emergency Response Use Case comes from several sources. There is emergency response data provided by the Sheffield City Council; posts from the Sheffield City Council forum; Flickr photos uploaded by Flickr users during emergency scenarios; and news articles covering emergency situations.

Dataset	Provider	User
2.1 Sheffield City Council Emergency Planning data	SCC	USFD, BUT
2.2 ABC News articles	USFD	USFD
2.3 Yahoo Geoplanet/OpenStreetMap Mapping	USFD	USFD
2.4 Flickr ER Sheffield	CERTH-ITI	CERTH-ITI

Table 3: Datasets used in the Emergency Response Use Case

2.1 Sheffield City Council Emergency Planning data

A wide variety of data is generated and collected during an emergency and SCC Emergency Planning has supplied a wide variety of datasets for the project partners, such as, event logs, police logs, fire service logs, CCTV images, CCTV videos, fire and rescue service phone calls. This data included text documents, images, audio and video. This collection is described in more detail in D6.5.1.

BUT employed the data set originated at the Fire Service of Sheffield. It was exported from the proprietary system used for call auditing. The data corresponds to recordings of two days during Sheffield flooding in 2007. It was delivered as WAV files in linear encoding. According to the available documentation of the recording device, the data is processed by the G729A codec. The data were processed by the recognition system and the accuracy was assessed (results summarized in D1.2.1).

Deliverables: D6.5.1, D1.2.1.

Publications:

F. Grezl, M. Karafiat, and L. Burget. Investigation into bottle-neck features for meeting speech recognition. In Proceedings of InterSpeech 2009. [7]

Services: wp2-speech-indexing, wp2-search-in-speech

2.2 ABC News articles

a. Tagged articles downloaded from ABC News archive site⁹. There are also image and audio (with transcripts) on the site which will be also uploaded onto the WebDAV server. The site is copyrighted so the data cannot be distributed or published. This first set of data relates to text articles, there are 3198 documents of the format shown in Figure 1.

⁹<http://www.abc.net.au/news/tag/>

```
<Document>
<URL>
<![CDATA[http://www.abc.net.au/news/stories/2009/08/31/2672352.htm?site=news]]>
</URL>
<Title><![CDATA[More swine flu deaths]]></Title>
<Summary>
<![CDATA[Three more people with swine flu have died in Western Australia.]]>
</Summary>
<Text><![CDATA[Three more people with swine flu have died in Western Australia.
A 52-year-old man died at home last week and a 64-year-old woman died at Royal
Perth hospital on Saturday. A 47-year-old man died at Sir Charles Gairdner
hospital on Saturday. It is believed he did not have any underlying medical
conditions. The deaths take the state's toll to 23. More than 4,000 people have
tested positive to the virus in WA.]]></Text>
<CreationTime><![CDATA[Mon Aug 31, 2009 7:22pm AEST]]></CreationTime>
<TagURI><![CDATA[http://www.abc.net.au/news/tag/health]]></TagURI>
<TagURI><![CDATA[http://www.abc.net.au/news/tag/swine-influenza]]></TagURI>
<TagURI><![CDATA[http://www.abc.net.au/news/tag/wa]]></TagURI>
<TagURI><![CDATA[http://www.abc.net.au/news/tag/perth-6000]]></TagURI>
</Document>
```

Figure 1: Example text article from the ABC News dataset

b. An enhanced version of this data, containing 6947 articles using an extended number of ER related tags to crawl the site. The tags are: health, road-accidents, accidents, disasters-and-accidents, maritime-accidents, storm, pollution, emergency-planning, residential-fires, industrial-fires, floods, infectious-diseases, emergency-incidents, avian-influenza, swine-influenza, rail-accidents, water-pollution, land-pollution, air-and-space-accidents, relief-and-aid-organisations, earthquake, tidal-wave, weather-phenomena, water-management, drought, bushfire, event, weather, air-pollution, electricity-energy-and-utilities, oil-and-gas, home-accidents, influenza, telecommunications, water-supply, workplace-accidents

Deliverables: N/A.

Publications: N/A.

Services: Currently, the tag graph "abcnews_tag_graph.tgr" used by the local tag community detection service has been created from this dataset. This data forms part of the modelling data used to construct the ER Tagging Model used by the Text Classification System (WP2.Text.Classification)

2.3 Yahoo Geoplanet/OpenStreetMap Mapping

The data provides (sameAs) links between the locations in the Yahoo Geoplanet data¹⁰ and the OpenStreetMap data¹¹. The text analysis geocoding system currently uses the Geoplanet version 7.5.2 (released 2010-06-30) and the OpenStreetMap database dump from 2010-09-15. For these releases the approximate number of locations are 5.3 million for Geoplanet and 785 million for OpenStreetMap. Not all Geoplanet location place types are mapped (Zones, TimeZones and Zip/Postcodes are ignored) of the remaining (approx 1.15 million locations), approximately 60% of the Geoplanet locations are assigned a matching OpenStreetMap location.

¹⁰<http://developer.yahoo.com/geo/geoplanet>

¹¹<http://planet.openstreetmap.org/>

Deliverables: N/A.

Publications: N/A.

Services: Used to provide geo-tagging in the Text Annotation Service (WP2_Text_Annotation).

2.4 Flickr ER Sheffield

136 ER-related Sheffield photos from Flickr. Two sets of Flickr photos that involve ER incidents in Sheffield that occurred in June 2007. These sets comprise 117 photos of Sheffield flood and 19 photos of the Sheffield gatecrasher fire (ER-Sheffield photos and metadata, June2007). This dataset resulted as an outcome of the WP2_SemanticPhotoQuery service.

Deliverables: The WP2_TagNormalization and WP2_SemanticPhotoQuery service that were used to produce this dataset are/will be described in D2.1.2 "Intelligent media analysis tools" and D2.3 "Social Media Intelligence techniques", respectively.

Publications: N/A.

Services: Used in the evaluation of WP2_TagNormalization and WP2_SemanticPhotoQuery.

3 Research Evaluation Datasets

Apart from the datasets used in the services underlying the two set of prototypes, a number of datasets have been created to evaluate various components of the services and for evaluating related research activities. These include several Flickr datasets used for evaluating visual analysis and tag analysis research; Twitter and Answerbag crawls for evaluating research on community structures and community membership lifecycle; and Delicious and Bibsonomy datasets for evaluating tagging system; and Wikipedia/DBpedia used for evaluating entity disambiguation.

A list of Research Evaluation datasets is given in Table 4

Dataset	Provider	User
3.1 FLICKR-1M	CERTH-ITI	CERTH-ITI
3.2 Flickr Datasets with visual features	CERTH-ITI	CERTH-ITI
3.3 Flickr dataset with temporal info	CERTH-ITI	CERTH-ITI
3.4 Flickr Logo database	Yahoo!	Yahoo!
3.5 Flickr Data for Sheffield/Cardiff/Cambridge areas	USFD	USFD
3.6 Sheffield City Council major incidents	SCC/USFD	USFD
3.7 Twitter community dataset	EMKA	EMKA
3.8 Communications structures	EMKA	EMKA
3.9 Wiki/DBpedia	UKob	UKob
3.10 BIBSONOMY-200K	CERTH-ITI	CERTH-ITI
3.11 DELICIOUS-7M	CERTH-ITI	CERTH-ITI

Table 4: Datasets used in the Research Evaluation

3.1 FLICKR-1M

Flickr is a popular online photo sharing and organizing application. For our experiments, we used a focused subset of Flickr comprising approximately 120,000 images that were located within the city of Barcelona (by use of a geo-query). In total, the number of tag assignments for this dataset approaches one million. The dataset contains only the tag assignments, i.e. entries of the form "user", "photo_id", "tag" and any associated metadata or content (the photos themselves) need to be retrieved by the dataset user.

Deliverables: D3.3

Publications:

S. Papadopoulos, Y. Kompatsiaris, A. Vakali. A Graph-based Clustering Scheme for Identifying Related Tags in Folksonomies. In Proceedings of DaWaK'10, 12th International Conference on Data Warehousing and Knowledge discovery. 2010 [17]

S. Papadopoulos, A. Vakali, Y. Kompatsiaris. Community Detection in Collaborative Tagging Systems. In Book Community-built Database: Research and Development. 2010 [15]

Services: This is very similar to the above dataset which is used in the tag recommendation service (See section 1.5).

3.2 Flickr Datasets with visual features

Two different datasets from Flickr were crawled using the wget utility and Flickr API facilities. The first one consists of 3000 images depicting cityscape, seaside, mountain, roadside, landscape, sport-side and locations (about 500 images from each domain). The second dataset comprises 20000 images related to concepts: jaguar, turkey, apple, bush, sea, city, vegetation, roadside, rock, tennis. As a source of semantic information for tag concepts, we employ the lexiconWordNet, which stores English words organized in hierarchies, depending on their cognitive meaning. Both image datasets were manually annotated in order to get the ground truth. The gathered dataset together with the manual annotations is a valuable source for the training of multimedia analysis algorithms. For each object in the two datasets, we have: the resource (jpg file), tags, MPEG-7 visual features of the resource.

Deliverables: To be included in D2.4 Media analysis evaluation

Publications:

E. Giannakidou, F. Kaklidou, E. Chatzilari, I. Kompatsiaris, A. Vakali. Harvesting Intelligence in Multimedia Social Tagging Systems. In *Emergent Web Intelligence: Advanced Information Retrieval*. 2010. [4]

E. Chatzilari, S. Nikolopoulos, E. Giannakidou and I. Kompatsiaris. Leveraging Social Media For Training Object Detectors. In *16th International Conference on Digital Signal Processing (DSP'09)*. 2009 [2]

S. Nikolopoulos, E. Chatzilari, E. Giannakidou and I. Kompatsiaris. Towards fully un-supervised methods for generating object detection classifiers using social data. In *10th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2009)* [13]

E. Giannakidou, I. Kompatsiaris, A. Vakali. SEMSOC: SEMantic, SOcial and Content-based Clustering in Multimedia Collaborative Tagging Systems. In *Proc. 2nd IEEE International Conference on Semantic Computing (ICSC' 2008)* [5]

Services: Expected to be used by WP2_GetContentbasedWeight service.

3.3 Flickr dataset with temporal info

A Flickr dataset with photos and their associated metadata (i.e. tags, uploading time, user) which were uploaded during the time period September 2007 - September 2008. After the preprocessing step, we resulted in a dataset of 1218 users, 6764 photos and 2496 unique tags that span in 210 days.

Deliverables: D3.4 Community browser and graphical user interface elements and evaluation

Publications:

E. Giannakidou, V. Koutsonikola, A. Vakali and I. Kompatsiaris. Exploring Temporal Aspects in User-Tag Co-Clustering. In *Proc. 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010)* [6]

V. Koutsonikola, A. Vakali, E. Giannakidou I. Kompatsiaris. Clustering Users of a Social Tagging System: A Topic and Time Based Approach. In *Proc. of the 10th International Conference*

on Web Information Systems Engineering (WISE 2009) [12]

Services: It is going to be used by a stand-alone application to visualize user and tag activity over time.

3.4 Flickr Logo database

To evaluate Logo Detection experiments a dataset has been built with the following features:

A total of 27 commercial brands was manually picked with the following criteria: (i) To be able to find enough test images in natural environments, (ii) to have a variety of topics (not just car brands, for instance). 40 images were manually selected per selected brand from Flickr, using Flickr Search with the name of the brand as query, ensuring that every selected image effectively contained the logo image.

Once the initial collection was built, all 1080 images were annotated with the bounding box coordinates of the logo in the image. The annotation application allowed for multiple bounding box definition per image, for the cases of images containing more than one logo instance.

This annotated collection of logos was then split in a test set and a training set. From the 40 images per brand, 30 were randomly selected to be part of the training set, while 10 were randomly selected for the test set. The test set was then randomly split in 6 subsets and the test set was again randomly split in 2 sets. At the end of the process, we had 6 groups of 5 images for training and 2 groups of 5 images for testing per brand.

The query set was then formed with one group of the test set (5 images times 27 brands = 135 query images). To complete this set, 135 images were manually picked with two criteria: (i) to ensure that they did not contain a logo and, (ii) to ensure that they were natural images as the ones previously selected.

Finally, the distractor set was built: 4397 images from Flickr were selected from the group "Identity + Logo Design". All images in this group contain logo images in a not natural environment. The bounding box annotation was not extracted, as it was assumed that the whole image represented the logo.

Deliverables: N/A.

Publications: A publication using this dataset is currently in writing but has not been submitted yet.

Services: N/A.

3.5 Flickr Data for Sheffield/Cardiff/Cambridge areas prior to 2010

Three location areas were chosen; Cambridge, including Ely, Newmarket and Haverhill (as a classical example of an ambiguous location); Sheffield, including, Chestereld, Barnsley, Hope Valley and Rotherham (for which an accurate local geographic database is available); and Cardiff, including Barry, Ferndale, Sully, Penarth, Porth, Bridgend, Aberdare, Mountain Ash, Pentre, Cowbridge (which offers a number of highly non-ambiguous location names, and location names which are ambiguous due to also being common terms, namely Barry, Sully and Mountain Ash).

These 20 location names can be resolved into 268 toponyms. In total 1143529 photos were tagged with at least one of these terms (after removing duplicates), of which 123124 (10.8%) have an associated geolocation (latitude/longitude), these were uploaded by 12326 users (approximately 10 photos per user).

The photos contain 165389 location name tags, note that each photo must contain at least one location tag to be retrieved. The users 580296 Flickr contacts produce 1140668 location name tags and the contacts 5700749 contacts produce 3998763 location name tags. Whilst all the collected data was limited to an upload date before the end of 2009, all the contact and tag values are up-to-date at the time of retrieve (March 2010).

Deliverables: The data was described in the contextual analysis experiments reported in D2.2 "Contextual media analysis and fusion techniques"

Publications:

N. Ireson and F. Ciravegna Toponym Resolution in Social Media To be published in the 9th International Semantic Web Conference. 2010. [9].

Services: N/A.

3.6 Sheffield City Council major incidents 2006-9

Data relating to the 65 major incidents reported by the Sheffield City Council between 2006-9, the data includes: incident date, location of the incident, a short textual description, a set of keywords and an estimation of incident severity by the members of the SCC team. The location and keyword information was then used as a query to retrieve potentially related information, both from the Internet in general and from the main public forum in the Sheffield area¹². The most major incident in this time-frame was the 2007 flooding for which a separate data set is available which was described and used in Deliverable 2.1.1 "Initial implementation of intelligent media analysis tools"

Deliverables: D2.1.1

Publications:

Part of the data set (that relating to the Sheffield 2007 Floods) was used in: N. Ireson Local Community Situation Awareness During an Emergency. Proceedings of the IEEE International Conference on Digital Ecosystems and Technologies (IEEE-DEST 2009) [8]

Services: N/A.

3.7 Twitter community dataset

Based on so-called hash-tags, e.g. #WeKnowIt, we detected communities on Twitter¹³. The Community Membership Life Cycle - mentioned in D4.2 - was applied on the data to identify social roles.

¹²<http://www.sheffieldforum.co.uk>

¹³<http://www.twitter.com>

For evaluating and testing the Community Membership Life Cycle Model developed in T4.2 Community Administration Platform we used a data set of Twitter data, collected by ourselves.

Deliverables: related with work done in D4.2

Publications:

A. Sonnenbichler, and C. Bazant. Application of a Community Membership Life Cycle Model on Tag-based Communities in Twitter. In Proceedings of the 34th Annual Conference of the German Classification Society (GfKI). 2010. [21]

A. Sonnenbichler. A Community Membership Life Cycle Model. 2010. [22]

Services: N/A.

3.8 Communications structures

For evaluating communications structures and clustering algorithms several public datasets were crawled. For evaluating the hierarchicalness of Q&A posting networks a dataset was crawled from Answerbag¹⁴ (like Yahoo Answers¹⁵). The dataset of Answerbag contains all questions and answers written from the forum opening on May 27th, 2003 up to November 3rd, 2008.

Deliverables: N/A.

Publications:

M. Ovelgönne. On the Hierarchicalness of Q&A Posting Networks. In GROUP '10: Proceedings of the ACM 2010 international conference on Supporting group work (to appear). 2010. [14]

Services: N/A.

3.9 Wiki/DBPedia

Wikipedia/DBPedia as source dataset + Reuters corpora of news for testing. Used for entity detection/disambiguation. Source dataset was extracted from Wikipedia/DBpedia to include all named entities (pages from Wikipedia), relationships between them (info-boxes, templates and simple hrefs) and all possible entity namings (alternative names, disambiguation, etc.). This dataset is used to recognize and categorize entities recognized in document. Categorization testing and verification is done using Reuters news corpora.

Deliverables: Part of services in D3.1 (answer quality measure) and D3.3 (named entity recognition and categorization)

Publications:

M. Janik and K. J. Kochut. OmniCat: Automatic Text Classification with Dynamically Defined Categories. In Poster Session at 7th International Semantic Web Conference (ISWC 2008). 2008. [10]

¹⁴<http://www.answerbag.com/>

¹⁵<http://answers.yahoo.com/>

Services: N/A.

3.10 BIBSONOMY-200K

BibSonomy is a social bookmarking and publication sharing application. The BibSonomy dataset was made available through the ECML PKDD Discovery Challenge 2009. We used the Post-Core version of the dataset, which consists of a little more than 200,000 tag assignments (triplets) and hence the label 200K was used to form the dataset name. The dataset is publicly available¹⁶.

Deliverables: D3.3 Mass Classification and Clustering Tools

Publications:

S. Papadopoulos, Y. Kompatsiaris, A. Vakali. A Graph-based Clustering Scheme for Identifying Related Tags in Folksonomies. In Proceedings of DaWaK'10, 12th International Conference on Data Warehousing and Knowledge discovery. 2010 [17]

S. Papadopoulos, A. Vakali, Y. Kompatsiaris. Community Detection in Collaborative Tagging Systems. In Book Community-built Database: Research and Development. 2010 [15]

S. Papadopoulos, Y. Kompatsiaris, A. Vakali. Leveraging Collective Intelligence through Community Detection in Tag Networks. In Proceedings of CKCaR'09 Workshop on Collective Knowledge Capturing and Representation. 2009. [16]

Services: Due to the theme of the dataset (scientific citations), there is no direct relation to any of the two scenarios. Thus, there is no plan to use the dataset in any of the services.

3.11 DELICIOUS-7M

Delicious is a popular social bookmarking service for managing and sharing bookmark collections. We used a snapshot of the Delicious bookmark collection corresponding to January 2006, comprising seven million tag assignments. This dataset is a subset of the collection studied in (Wetzker et al., 2008)¹⁷. The dataset only contains quadruples of the form "user", "url", "tag", "date", and sometimes the tag is missing (sometimes users only bookmark a URL and do not add any tags to it).

Deliverables: D3.3 Mass Classification and Clustering Tools

Publications:

S. Papadopoulos, Y. Kompatsiaris, A. Vakali. A Graph-based Clustering Scheme for Identifying Related Tags in Folksonomies. In Proceedings of DaWaK'10, 12th International Conference on Data Warehousing and Knowledge discovery. 2010 [17]

S. Papadopoulos, A. Vakali, Y. Kompatsiaris. Community Detection in Collaborative Tagging Systems. In Book Community-built Database: Research and Development. 2010 [15]

¹⁶<http://www.kde.cs.uni-kassel.de/bibsonomy/dumps>

¹⁷http://www.dai-labor.de/en/competence_centers/information_retrieval_machine_learning/datasets

Services: Due to the temporal restriction of the dataset (bookmarks in January 2006), there is no direct relation to any of the two scenarios. Thus, there is no plan to use the dataset in any of the services.

4 Data exchange

The collection of datasets and their use as input in various services was performed by individual partners. In case two or more partners collaborated in the use of a dataset (see e.g. 1.1, 2.2, etc.) the dataset was loaded to as WebDAV server where all partners could access.

Some datasets had restricted copyright (e.g., the Lycos and Yahoo! Answers datasets described in D6.5.1) or contained sensitive information (e.g., Sheffield City Council Emergency Planning data in Section 2.1). In those cases the data was provided by the data owners to the individual partners on a case by case basis where the partners signed a restricted license agreement.

The combination of datasets was done by the service providers where appropriate but also in the two use case prototypes by fusing results from multiple services.

5 Conclusions

In this deliverable we have given an overview of the datasets used in the WeKnowIt project. The datasets are heterogeneous and include text, tags, images, audio, video, entities, gazetteers, communication and community structures. The goal of the deliverable has been to give a complete, yet shallow, overview of the datasets used in the project. For more in-depth descriptions of the datasets and their usage we refer to the corresponding WeKnowIt deliverables and other publications referenced in this document.

6 References

- [1] Y. Avrithis, Y. Kalantidis, G. Toliás, and E. Spyrou. Retrieving landmark and non-landmark images from community photo collections. In *in Proceedings of ACM Multimedia (MM 2010)*, Firenze, Italy, October 2010.
- [2] Elisavet Chatzilari, Spiros Nikolopoulos, Eirini Giannakidou, and Ioannis Kompatsiaris. Leveraging social media for training object detectors. In *DSP'09: Proceedings of the 16th international conference on Digital Signal Processing*, pages 65–76. IEEE Press, 2009.
- [3] S. Diplaris, A. Flores, B. Sigurbjörnsson, N. Tintarev, M. Escriche, R. van Zwol, and Y. Kompatsiaris. Collective intelligence in mobile consumer social applications. In *9th International Conference on Mobile Business (ICMB 2010)*, pages 206 – 212, 2010.
- [4] Eirini Giannakidou, Fotini Kaklidou, Ioannis Kompatsiaris, and Athena Vakali. Harvesting intelligence in multimedia social tagging systems. In Richard Chbeir, Youakim Badr, Ajith Abraham, and Aboul-Ella Hassanien, editors, *Emergent Web Intelligence: Advanced Information Retrieval*, Advanced Information and Knowledge Processing, pages 135–167. Springer London, 2010.
- [5] Eirini Giannakidou, Ioannis Kompatsiaris, and Athena Vakali. Semsoc: Semantic, social and content-based clustering in multimedia collaborative tagging systems. In *ICSC '08: Proceedings of the 2008 IEEE International Conference on Semantic Computing*, pages 128–135, Washington, DC, USA, 2008. IEEE Computer Society.
- [6] Eirini Giannakidou, Vassiliki Koutsonikola, Athena Vakali, and Yiannis Kompatsiaris. Exploring temporal aspects in user-tag co-clustering. In *Proc. 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2010)*, 2010.
- [7] Frantisek Grezl, Martin Karafiat, and Lukas Burget. Investigation into bottle-neck features for meeting speech recognition. In *Proceedings of InterSpeech 2009*, 2009.
- [8] Neil Ireson. Local community situation awareness during an emergency. In *Proc. of 2009 Third IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2009)*, Istanbul, Turkey, May 2009.
- [9] Neil Ireson and Fabio Ciravegna. Toponym resolution in social media. In *Proceedings of the 9th International Semantic Web Conference (ISWC 2010)*., 2010.
- [10] Maciej Janik and Krys J. Kochut. OmniCat: Automatic Text Classification with Dynamically Defined Categories. In *Poster Session at 7th International Semantic Web Conference (ISWC 2008)*, 2008.
- [11] Y. Kalantidis, G. Toliás, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, and S. Kollias. Viral: Visual image retrieval and localization. *Multimedia Tools and Applications (Submitted)*, 2010.
- [12] Vassiliki Koutsonikola, Athena Vakali, Eirini Giannakidou, and Ioannis Kompatsiaris. Clustering of social tagging system users: A topic and time based approach. In Gottfried Vossen, Darrell Long, and Jeffrey Yu, editors, *Web Information Systems Engineering - WISE 2009*, volume 5802 of *Lecture Notes in Computer Science*, pages 75–86. Springer Berlin / Heidelberg, 2009.

- [13] Spiros Nikolopoulos, Elisavet Chatzilari, Eirini Giannakidou, and Ioannis Kompatsiaris. Towards fully un-supervised methods for generating object detection classifiers using social data. In *WIAMIS '09: 10th Workshop on Image Analysis for Multimedia Interactive Services*, pages 230–233, 2009.
- [14] Michael Ovelgönne. On the hierarchicalness of q&a posting networks. In *GROUP '10: Proceedings of the ACM 2010 international conference on Supporting group work (to appear)*, New York, NY, USA, 2010. ACM.
- [15] Symeon Papadopoulos, , Athena Vakali, and Yiannis Kompatsiaris. Community detection in collaborative tagging systems. In *Book: Community-built Database: Research and Development*, Springer, 2010.
- [16] Symeon Papadopoulos, Yiannis Kompatsiaris, and Athena Vakali. Leveraging collective intelligence through community detection in tag networks. In *CKCaR '09: Workshop on Collective Knowledge Capturing and Representation*, 2009.
- [17] Symeon Papadopoulos, Yiannis Kompatsiaris, and Athena Vakali. A graph-based clustering scheme for identifying related tags in folksonomies. In *DaWaK '10: Proceedings of the 12th International Conference in Data Warehousing and Knowledge Discovery*, pages 65–76. Springer-Verlag, 2010.
- [18] Symeon Papadopoulos, Christos Zigkolis, Stefanos Kapiris, Yiannis Kompatsiaris, and Athena Vakali. Clusttour: City exploration by use of hybrid photo clustering. In *MM '10: Proceedings of the ACM International Conference in Multimedia*, 2010.
- [19] Symeon Papadopoulos, Christos Zigkolis, Giorgos Tolia, Yannis Kalantidis, Phivos Mylonas, Yiannis Kompatsiaris, and Athena Vakali. Image clustering through community detection on hybrid image similarity graphs. In *ICIP '10: Proceedings of the 2010 International Conference in Image Processing*. IEEE, 2010.
- [20] Sergej Sizov and Andreas Ens. Eventfolk - automatic event detection in social media. *Datenbank-Spektrum*, 10(1):7–13, 2010.
- [21] Andreas Sonnenbichler and Christopher Bazant. Application of a community membership life cycle model on tag-based communities in twitter. In Karlsruhe Institute of Technology, editor, *Proceedings of the 34th Annual Conference of the German Classification Society (GfKI)*, Karlsruhe, Germany, 2010.
- [22] Andreas C Sonnenbichler. A community membership life cycle model. *1006.4271*, June 2010.
- [23] Roelof van Zwol, Lluís Garcia Pueyo, Mridul Muralidharan, and Börkur Sigurbjörnsson. Machine learned ranking of entity facets. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 879–880, New York, NY, USA, 2010. ACM.
- [24] Roelof van Zwol, Börkur Sigurbjörnsson, Ramu Adapala, Lluís Garcia Pueyo, Abhinav Katiyar, Kaushal Kurapati, Mridul Muralidharan, Sudar Muthu, Vanessa Murdock, Polly Ng, Anand Ramani, Anuj Sahai, Sriram Thiru Sathish, Hari Vasudev, and Upendra Vuyyuru. Faceted exploration of image search results. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 961–970, New York, NY, USA, 2010. ACM.