www.forgetit-project.eu

ForgetIT

Concise Preservation by Combining Managed Forgetting
and Contextualized Remembering

Grant Agreement No. 600826

Deliverable D6.1

Work-package	WP6: Contextualization / Decontextualization
Deliverable	D6.1: State of the Art and Approach for Contextualization
Deliverable Leader	Mark A. Greenwood
Quality Assessor	Heiko Maus
Estimation of PM spent	6
Dissemination level	Public
Delivery date in Annex I	31st July 2013
Actual delivery date	17th July 2013
Revisions	3
Status	Final
Keywords:	context, text, images, ontologies, preservation

Disclaimer

This document contains material, which is under copyright of individual or several ForgetIT consortium parties, and no copying or distributing, in any form or by any means, is allowed without the prior written agreement of the owner of the property rights.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the ForgetIT consortium as a whole, nor individual parties of the ForgetIT consortium warrant that the information contained in this document is suitable for use, nor that the use of the information is free from risk, and accepts no liability for loss or damage suffered by any person using this information.

This document reflects only the authors' view. The European Community is not liable for any use that may be made of the information contained herein.

List of Authors

Andrea Ceroni (LUH)

Mark A. Greenwood (USFD)

Nattiya Kanhabua (LUH)

Vasileios Mezaris (CERTH)

Claudia Niederee (LUH)

Jörgen Nilsson (LTU)

Olga Papadopoulou (CERTH)

Contents

Executive summary

Whilst documents, photos, and videos are useful mediums for telling us about the past as time goes by they become open to interpretation as their context is lost or changes beyond recognition. Although the ForgetIT project is not dealing with millennial old historical documents, it is aiming to support the long-term archival storage of newly created documents. While we can easily interpret documents we create today, there is no guarantee that even the author will be able to fully interpret them in a year or a decade from now, and for personal preservation following generations may struggle to interpret documents written by their forefathers. The ForgetIT project aims to reduce the burden of interpreting archived documents by providing contextualization components that collect context at archival time and then re-contextualize documents as they are retrieved from the archive at a later date.

The rest of this deliverable is divided into four main sections. Firstly, in Section 1, we define exactly what we mean by *context*, before we outline the approach to contextualization that we intend to follow within the ForgetIT project. Section 3 reports a review of the state-of-the-art for the techniques we intend to utilize. Finally, Section 4 details our expected research plan for the next six months and the first release of tools for contextualization, which will be reported in deliverable D6.2.

It is worth noting that the techniques and planned work outlined in this deliverable are only applicable to those documents freely selected for preservation. Those documents which have to be retained for regulatory purposes are outside the scope of this work as legally they have to be stored as-is without alteration or addition.

1 What Is Context?

> **contextualize**, *v.* To place in, or treat as part of, a context [1]

Adequate context is vital for the correct interpretation of stored information. Just as words in isolation are often ambiguous (is a mention of 'apple' referring to the fruit or the consumer electronics company?) so textual documents, photos and videos are enhanced by knowledge of their context. Imagine flicking through the holiday photos of a total stranger; the chances are that you would be unable to form a coherent and correct understanding of their travels. In a similar way a single document from a company website is likely to be open to misinterpretation without access to either the rest of the site, or possibly internal company documents.

One of the most challenging aspects of dealing with context is that it is never explicitly stated; a considerable portion of the context required interpretation residies within the memory of those doing the interpretation. This might be general background knowledge, expert knowledge, knowledge about processes in a company, or personal knowledge about a holiday drip. In addition relevant context might be distributed over various items of information such as documents, photos, process and structure descriptions, etc.

If an information object has to be interpreted in a context, which is different from its context of creation, some of the information about its original context has to made explicit and associated with the information object, in order to ensure a consistent interpretation. This is the process of *contextualization*. In ForgetIT, we are specifically interested in being able to ensure that an information object can be fully understood many years into the future. This can be formally stated as:

> For a piece of information i, **contextualization** is the process of providing sufficient additional information c^{+} (context information) such that i can be interpreted/understood in a future context of interpretation C^{F} in a similar way as it has been interpreted in the original context C^{O}:
>
> $$I(i, C^{O}) \cong I((i \oplus c^{+}), C^{F}) \tag{1.1}$$
>
> where the function $I(i, C)$ refers to the interpretation of the information object i in context C.

The process of contextualization is non-trivial, and requires at least some knowledge about the intended interpretation in the original context as well as an estimation, of what will be required in order to understand the information object in any future context.

For long-term archiving, as envisaged by the ForgetIT project, where there could easily be decades between C^{O} and C^{F}, it becomes vitally important that adequate context is preserved along with each item that is stored. Such context, allows archived items to be fully and correctly interpreted at some undefined future date. This leaves us needing to answer a number of important questions; *what are relevant aspects of context in the context of preservation and for different types of information objects? how are archived items linked to their context?*, *how should context (information) be stored?* and *how do we keep the context information up to date?* Answers to these questions form the backbone of the rest of this deliverable, while the remainder of this section aims to give a brief overview of exactly what we mean by context within the scope of the ForgetIT project.

The context of a document can be defined from two alternative perspectives. Firstly we have the explicit information which *describes* the document; when and by whom the document was authored, where a photo was taken, etc. The context of a document can also be defined as the set of information required to *understand* a document after a period of time has elapsed. When considering long-term archiving this period could easily be 10, 20, or more years.

A nice metaphor for context is to think of the information to be in a box, which forms the context of its interpretation [2]. This box is described by a set of parameters such as time, location, language interpretation, which are left implicit in the content of the box, since they are obvious or implicitly known. As a simple example think of the sentence ”It's raining today”. It for example does not say anything about the location, since it is implicitly known between the speakers. When we are now moving information, such as the above sentence, to another box (i.e. context), some of the parameters of the original box have to be made explicit depending upon the parameters of the original box. In ForgetIT, we are especially focussing on context transition with large gaps of time between the boxes (implied by long-term preservation).

There has been a lot of discussion on context dimensions in previous work, especially in the context of formal context models. A set of 12 context dimensions has, for example, be propagated by CYC [3]. Our initial set of context dimensions is inspired by this earlier work as well as by the special requirements of using the context information for contextualization in a preservation context. In identifying relevant context dimensions, we have focussed on two scenarios; personal information management and organizational information management.

In general, context can be described along a wide variety of dimensions. The initial challenge is to identify a set of dimensions of context together with features to be considered in the individual context dimensions, which are useful for a long-term preservation setting. Based on this starting point, a subset of features will be selected depending on the type and granularity of document to be contextualized, the concrete preservation setting (personal vs. organizational preservation) and other factors.

The set of context dimensions and associated features currently under consideration are:

- **Time:** The temporal dimension is a crucial component of context and can be subdivided as follows:
  - *Creation Time:* the time at which the document has been created.
  - *Content Time:* this is the time the content refers to and which can, for example, be encoded as temporal expressions inside the document (e.g. yesterday, last

week, etc.).

- **Location:** This refers to locations associated with an information object. For photos this is likely to be the location the photo was taken, whereas a textual document may well mention numerous locations which may relate to the context of the document (e.g. the location of a holiday, the headquarters of the company, etc.). It is also important to take into account the hierarchical nature of location information (i.e. Sheffield is a town in South Yorkshire, which is part of England, which is part of the UK, which is within Europe) to allow for aggregation and spatial search [4].

- **Topic:** The topic(s) of the document or document section. As the meaning of a topic might change over time, it will be important to identify the specific interpretation commonly assumed at the point of archiving the document.

- **Entity Space:** This refers to the entities associated with the information object under consideration. This might be different types of entities such as persons, organizations, events, etc. Features to be considered here are named entities contained in a document, persons or monuments depicted on a photo, etc.

- **Document Space:** Information objects do not usually occur in isolation, but as part of a logical collection of some form (a company website, photos from a single trip, etc.). Those related documents contribute to the understanding of the information object under consideration.

Together these context dimensions should provide enough information to allow us to both describe and understand any document (text, image, video, etc.) we wish to archive and as such will form the backbone of the context archived with them as well as the context we will aim to re-create on retrieving an archived item. The rest of this deliverable describes the approach we intend to follow to generating such context within the ForgetIT project as well as the relevant state-of-the-art.

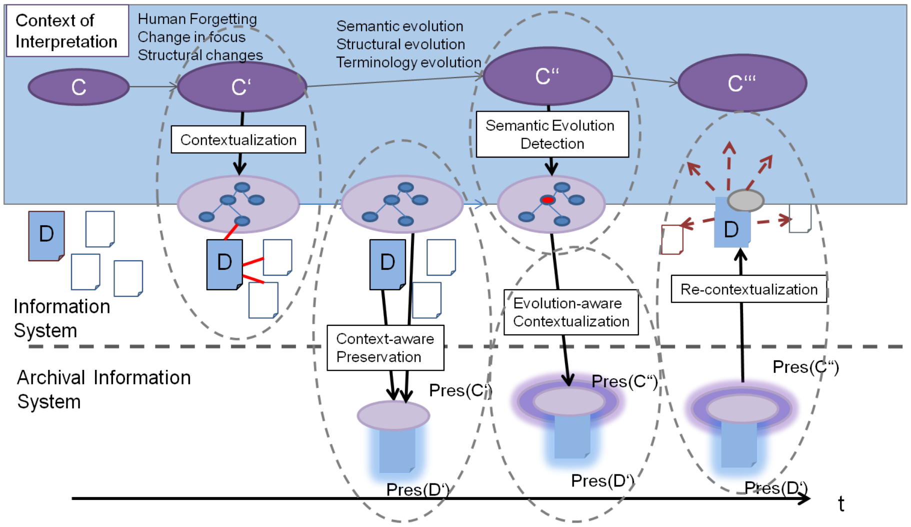

Figure 1: Contextualization Challenges in Preservation.

2 The ForgetIT Approach to Contextualization

As we have already show contextualization is a vitally important part of a long-term preservation strategy and our approach to the problem is summarized in Figure 1. This approach consists of the following four components:

- **Contextualization:** All documents are associated with context (this may be document metadata or external world knowledge, see Section 1) and this component is responsible for making the context explicit. We envisage that in most cases context will be encoded within an ontology. An ontology based context will bring multiple benefits and allow for a level of reasoning over the context (i.e. documents that mention Sheffield are also relevant if the user is interested in documents about the UK). We plan on utilizing existing techniques for information extraction (from both text and images) and disambiguation to link the extracted entities to an ontology. The exact ontologies used will depend on the specific usage scenario, but we would envisage using public ontologies (i.e. DBpedia) where possible.

- **Context-Aware Preservation:** This component is responsible for ensuring that when a document (textual or otherwise) is selected for preservation that the appropriate context is also archived. A naïve approach would be to simply archive a fresh copy of the entire context ontology with each item selected for preservation. While this would ensure that no context was lost it would be very wasteful. Many items will have the same context (e.g. a diary and a photo collection from a single trip would have the same context and company documents would all be written in the context of the same management structure) and as such the context should only be archived once.

- **Evolution-Aware Contextualization:** The ForgetIT project is focused on long-term digital preservation and over any reasonable time period the context of an individual document is likely to change and modelling this must be considered as part of the archiving strategy. Common changes in context would include changes in organizational roles, personal relationships, as well as unrelated world knowledge which may form part of a documents context (e.g. geo-political changes). In some situations simply archiving an updated version of the ontology (alongside the original) may suffice to model changes in context. In the worst case documents may need to be processed in order to understand how changes in context will affect their meaning; this would be an ideal job for a storlet working within the preservation archive.

- **Re-Contextualization:** The final step in the process of contextualization occurs when a document is retrieved from the archive at some future date. We envisage this being a user driven process with documents retrieved based on searching for specific items or via browsing the context of other documents (which may or may not be in the archive). Firstly it is important that once retrieved from the archive and put back into active use that the context is correctly retrieved and updated where necessary (possibly requiring techniques similar to those used for the initial contextualization), but equally important is that the process of re-contextualization should, where possible, not rely on a specific information management system, or even the ForgetIT framework as it would be presumptuous beyond measure to assume that the ForgetIT framework will remain unchanged (or will even still exist) within a century, yet we will still want to be able to retrieve items we preserve using the framework.

This overview, however brief, should enable the reader to visualize the process of contextualization and the components that will be created by the project. It is likely that numerous versions of each component will be developed either to target different document types (text, images etc.) of for different domains (personal versus organizational) as well as to experiment with different algorithms. The following section, Section 3, discuses the relevant state-of-the-art techniques which we intend to investigate and to develop to provide contextualization within the ForgetIT framework.

3 The Current State-of-the-Art

3.1 Information Contextualization

Whilst the context of textual documents may well be the same as that for images and videos it is clear that methods for determining the context will differ widely. As such this section is divided into two sub-sections. Firstly the relevant state-of-the-art algorithms for extracting context from textual documents is presented and then those for multimedia content.

3.1.1 Textual Content

Topics of interest for text analysis to perform contextualization include, but are not limited to, the following:

- Ontology Based Information Extraction (OBIE) techniques will be used to semantically annotate documents. The specific ontologies used will depend upon the use-cases. It is likely that standard Linked Open Data (LOD) resources will be useful for the organizational scenario (WP10), although it is less clear how applicable they will be to the personal preservation scenario (WP9).
- Ontologies, certainly in the personal preservation scenario (WP9), may not contain every important entity. In such situations it would be desirable to allow users to update their personal ontology (i.e. their PIMO, see D9.1 for more details). Without such updates re-contextualization may be difficult if not impossible. Such functionality will need to be developed within this workpackage but incorporated into the ForgetIT framework as being developed by WP8.
- The expansion, with relevant extra content, of image labels.

Identification of Named Entities (NE) such as people, organisations and locations is fundamental to semantic annotation and is the starting point of more advanced text mining algorithms. For instance, sentiment analysis is widely used in finance to extract the latest signals and events from news that could affect stock prices. However, before extracting company-related sentiment, it is necessary to identify the documents containing the corresponding and *unambiguous* company entities. Humans usually resolve ambiguities based on context and the current trend is to use publicly available Linked Open Data (LOD) to extend the context. While these techniques are currently focused on automatic approaches to entity disambiguation, they can easily be re-cast as approaches to contextualization of textual content. A good example of how this might work within ForgetIT is to look at how GATE Mímir [5, 6] can be used to combine external context with document content to provide a rich search experience.

GATE Mímir can be used to combine text, semantic annotations, and external knowledge bases into a single searchable index. The semantic annotations can be produced using any GATE application [7] allowing for a very flexible framework. As an example of how powerful such a combination can be, consider searching a collection of news articles for mentions of any member of the UK Labour party who also attended Edinburgh University being quoted. This is clearly not a query you could carry out using standard keyword search without a lot of prior research to determine the set of relevant people. Using GATE Mímir, however, the relevant documents can be located using the following query[1]:

```
{Person sparql = "SELECT ?inst WHERE {
       ?inst :party <http://dbpedia.org/resource/Labour_Party_%28UK%29> .
       ?inst :almaMater <http://dbpedia.org/resource/University_of_Edinburgh>
   }"
} [0..3] root:say
```

Whilst a full explanation of the GATE Mímir query syntax is beyond the scope of this discussion (see [5] for the full details) this query breaks down into three main parts.

- The query starts by looking for any mention of a Person within the documents. This is possible as the GATE application used to annotate the documents [8] creates `Person` annotations that span each mention of a person.
- Each `Person` annotation, where possible, has also been linked against an instance in DBpedia[2]. This allows us to restrict the set of `Person` annotations our query matches to just those which satisfy a SPARQL query; in this instance those people who are both a member of the UK Labour party and who were educated at the University of Edinburgh.
- The `Person` annotations which match the SPARQL query are then combined with a search of the document content to find those mentions which occur no more than three tokens before an instance of the verb to say (by searching on the root form of a token we will match say as well as saying, says, and said).

This query could be extended further to include document metadata (such as publication date etc.) but this simple example should hopefully illustrate the fact that the query is returning documents which do not explicitly mention the requested information; none of the documents returned mention the University of Edinburgh. In other words the documents are being retrieved based upon their context as well as their content. While search is not the focus of this work package it is clear that these techniques involved determining the context of the entities within the document and similar techniques will be explored as part of the contextualization work of ForgetIT.

In the field of Information Retrieval, *contextualization* has been defined as the process of estimating the relevance of a given document unit or a structural text by exploiting information coming from the surrounding document units or structural text [9]. Such definition is used in [10] as a starting point to perform contextualization of hyper-linked and semi-structured documents. In that work an *internal* and an *external* context are associated to every document. The former consists of the set of *incoming* and *outgoing* links in the

[1] You can try this query at `http://demos.gate.ac.uk/mimir/gpd/search/index`

[2] `http://dbpedia.org`

citation graph involving the document; the latter is represented by the internal and hierarchical structure of the document. The approach is then evaluated on the set of XML documents representing Wikipedia pages.

Spatial and temporal contextualization addresses the problem of defining the information context by taking into account both the spatial and temporal dimensions. Some research effort has been spent in this direction by the NEVAC[3], which contextualized unstructured documents by extracting implicit geographical and temporal references from them. As reported in [11], this was achieved through named entity recognition, relationship extraction and geographic information retrieval. As a further reference, an approach that focuses only on spatial contextualization, exploiting ontologies, can be found in [12].

Contextify[4] is an off-the-shelf product for email contextualization [13]. The context of a given email includes related emails, people, attachments and web links. Searching and visualization features are available to quickly filter and highlight search results. In addition to the list of conversation threads that match a given query, the network of people participating in those conversations is displayed as well. Each node in the network represents a person and the directed edges between the nodes tell who send emails to whom.

3.1.2 Multimedia Content

Multimedia content can provide useful information for image and video contextualization. Image similarity assessment and concept detection are two methods of multimedia analysis which can provide this information. As far as similarity is concerned, given a query image, several relevant images can be retrieved, and since the retrieved images are close to it, the information provided by them can be used for putting the original query image in context. Below we will briefly discuss techniques for similarity assessment between individual images and then we will extend these techniques for similarity assessment between collections of images. Moreover, we will present techniques for augmenting one image collection with images included in other similar collections providing more information to the image collection of interest. As far as concept detection is concerned, this can used for enhancing the assessment of image similarity that is calculated from low - level features alone.

Similarity assessment and concept detection for individual items

Many methods for similarity assessment have been introduced during the years trying to support the identification of similar or dissimilar multimedia items. First, in order to describe an image or video, a feature extraction technique is needed and several such techniques have been introduced. For feature extraction, the image characteristics that are used are either global image characteristics (global descriptors) or local image characteristics (local descriptors). After extracting a feature vector for each of the two or more

[3] North–East Visualization and Analytics Center

[4] http://contextify.net

media items that we want to compare, a similarity measure is used for deciding whether two items are similar to each other or not. For this, a number of description distance functions have been introduced as similarity measures, from the simple Euclidean distance to more specialized functions. For a more detailed presentation of the various feature extraction and similarity assessment techniques that have been proposed, we refer the reader to deliverable D4.1 of WP4.

As mentioned above, concept detection is another multimedia analysis method that can provide more information about the multimedia items and enhance the similarity assessment process. Concept detection typically involves the following steps: content sampling (e.g selection of a low - resolution version of the image or selection of keyframe of the video), to reduce the amount of visual information that will need to be processed; feature extraction, as described above; application of trained concept detectors (often based on Support Vector Machine classifiers or other machine learning techniques), to associate the low - level features extracted from the media item with one or more high - level concepts, e.g. 'outdoors', 'building' etc. that can describe the visual content in a human - understandable way. Then, these high level concepts can be used to complement the low - level features, for assessing the similarity between a pair of images. Again, we refer the reader to D4.1 for a detailed state-of-the-art report on concept detection.

Similarity assessment for image or video collections

We describe above how similarity assessment can provide useful information for image contextualization. An extension of it, i.e. a method for similarity assessment between image collections, rather that individual images, is even of greater importance in supporting this goal. Due to the increasing amount of multimedia items captured by various different users, many image collections end up containing images that are similar to those of one or more other collections, being in fact about the same real - life event, e.g. a specific music concert or sporting event. Whilst the collections exhibit both thematic and visual similarity, the information contained in each of them (e.g. some images that are in only one of these collections) or provided for each of them (e.g. any textual description of them or accompanying metadata) may differ a lot. For this, finding sets of images with similar content gives us the opportunity to augment the one which is lacking information in some respect, thus facilitating its correct interpretation. A simple way of comparing an image collection to another collection is to compare each image of the one collection to all images of the other independently. Reviewing the state-of-the-art in this area we only find one approach dealing with this challenge. As the authors in [14] report, they are the first who proposed a technique for similarity assessment between sets of images. Specifically, they introduced a new visual-based method for retrieving events in photo collections. In their work, each event is described by a set of images. The method is similar to object detection, and the authors make an analogy between the components they use in their method and the components of a typical object detection pipeline: event-records correspond to images; global visual features describing each picture of the record correspond to the local visual features describing points or regions of interest in an image; geo-coordinates and time stamps of the images correspond to spatial positions of the local features

In their approach, the authors make use of both visual content and contextual metadata. The information that is used for the similarity assessment between the sets of photos is time information, geographic information and visual content information. Time and geographic information are commonly used in retrieving events but have limitations which can be overcome by combining them with the visual content information. Thus, visual content is used at the first stage of their method to detect potential matches and the time and geo information is used at the second stage of the procedure in order to re - rank the results. Summarizing the method of [14], all images are associated with their geo - coordinate and time stamps and their visual content is described by a visual feature vector. Then the below steps are followed:

- Visual Matching: each query image feature is matched to the full features dataset. The matching is achieved with a distributed similarity search framework based on Multi-Probe Locality Sensitive Hashing and the MapReduce programming model [15, 16].
- Stop List: the records are filtered and those that have at least two matches are kept for the next steps.
- Geo-temporal consistency: a translation model between the query record and the retrieved records is computed. A final score is computed taking into account both geo and time metadata.
- Prior constraints: Depending on the application context, prior constraints are obtained on the acceptable values of the geo and time informations.

Evaluation of the proposed approach has been tested on LastFM-Flickr dataset showing good results.

As mentioned above similarity between image collections is performed in order to find collections with relevant characteristics and enhance the information of the collection of interest. Based on this, given one image collection we can try to find related images which are included in other similar collections and augment the collection of interest with that images. To our knowledge, there is no such approach in the state of the art. Nevertheless, we expect that it will be possible to accomplish this by using clustering algorithms. We believe that by dividing images into clusters it will be possible to find several images that will provide additional information and use them for augmenting the target collection. As for the clustering algorithms that may be used, several of them have been presented in deliverable D4.1 of WP4, and we refer the reader to it for more details.

Additionally, all the aforementioned techniques can be used either for image or video collection or for mixed collections.

While managing personal collections of multimedia documents, issues come from the great difference between multimedia objects and the high number of dimensions that are required to describe them. An approach to ontology–based multimedia document management, covering the semantic description of multimedia objects during their life–cycle, has been presented in [17]. The metadata of multimedia documents are extracted and then extended with both semantic and context information. The former is retrieved by querying an ontology, the latter is inferred from tracking user activities over time. Changes of such additional information are managed by taking into account the different phases of the document life–cycle.

3.2 Detecting Evolving Semantics

Accessing or searching archived collections (such as web archives) can be affected by terminology evolution over time [18], for instance, changes of words related to their definitions, semantics, and names (people, location, etc.). It is important to note that terminology evolution is a continuous process that can be observable also in a short term period caused by two major problems: 1) spelling variation in the modern and historic language, and 2) semantics changes over time (new words are introduced, others disappears, or the meaning of words changes). Previous work [19, 20] addressed the spelling variation problem using techniques from cross language information retrieval. They used probabilistic rule-based approaches for handling term variants when searching historic texts. In this case, a user can search using queries in contemporary language and the issued queries are translated into an old spelling possibly unknown to the user, which is similar to a query expansion technique in IR. Ernst-Gerlach and Fuhr proposed two ways to perform query expansion: an expansion of query and an expansion of index. In the first case, a set of rules is automatically constructed for mapping historic terms into modern terms. In the latter case, based on a lexical database, terms are indexed together with their synonyms and holonyms as additional indices.

Preliminary studies on the effect of named entity evolution in searching web archives are presented in [21, 22, 23, 24]. Berberich et al. proposed a method based on a hidden Markov model for reformulating a query to use time specific terminology. Kaluarachchi et al. studied the problem of concepts (or entities) whose names can change over time. They proposed to discover concepts that evolve over time using association rule mining, and used the discovered concepts to translate time-sensitive queries and answered appropriately. Tahmasebi et al. proposed to automatically detect terminology evolution within large, historic document collections by using clustering techniques and analysing co-occurrence graph. Kanhabua and Nørvåg defined a time-based synonym as a term semantically related to a named entity at a particular time period. They extracted synonyms of named entities from link anchor texts in Wikipedia articles using the full history and conducted experiments by measuring increased precision and recall in search results when employing time-based synonyms. The limitation of this approach is that an external knowledge source like Wikipedia does not cover all entities and are not able to capture ephemeral names or jargon used in everyday language or social media. In addition, none of previous work has addressed entity evolution in the context of personal and organizational archives as we will conduct in this project.

3.3 Context-Aware Preservation

Preservation is usually a deliberate undertaking, carried out by people determined and/or assigned to the task of preserving information. Therefore the *context-aware* part of the preservation is usually cared for manually by those who undertake this task (e.g. an archivist). The context as such is described by adding metadata to the preserved object, which in OAIS terminology [25] might be of either *context* or *provenance* type.

The archivists (or similar) usually also make the decisions on what to preserve, and what should be included as part of the object and as metadata. In this project, the idea is however that the ForgetIT system should assist in making those decisions, or even taking the decisions itself. If the system notices that some information has not been actively used for some time, it might decide to transfer it to the Archival Information System (AIS) by first transforming it into an Submission Information Package (SIP) and then initiate a transfer to the AIS. Preservation workflows are described in more detail in deliverable D5.1.

3.3.1 Defining the SIP

It might sound like a trivial task to determine what should be included in a Submission Information Package, and in its simplest form it probably is. You could certainly always send in everything that is contextually relevant for the object you want to preserve, including e.g. information about the creator and the creating organisation, all relevant contextual documentation (in other words, other information objects), the file format specifications for both the information object and for the metadata, and so on - but as you might realise, that would be quite an undertaking, and also most likely a waste of (some) effort since some of the information already might exist in the AIS. Typical information that already might exist is (not exclusive): file format specifications; creator information; organisation information; contextual information objects (ingested earlier); descriptions of software used for access

3.3.2 When to Transfer

Decisions on when to transfer information objects to an archive (or to preservation systems) is usually taken on a policy basis, e.g. 'this particular document type should be sent to the archive as soon as it is marked as closed and then retained for 10 years'. This could of course be handled automatically (e.g. by a document management system) as it is based upon well formulated rules.

The idea in the ForgetIT project is to make these decisions in a more fluid (but still automated) way, where some policies may exist but where the decisions are based on e.g. the usage frequency, age of object, date of relevance, related individuals, and so on.

4 Contextualization Research Plan

An overall high-level research plan as well as the goals for the research in contextualization has already been defined in the description of work (see especially the description of WP6 in the DoW). This section does not attempt to repeat that discussion but rather sets out the the research activities planned for the coming months leading up to month 12 and the first release of components for contextualization. The main focus of the next six months will be the following three research topics:

- formally defining a model of context to provide a theoretical underpinning to the components developed within the work package.
- prototype components that implement the four main strands of evolution-aware contextualization; context extraction, context aware preservation, context evolution, and re-contextualization (see Section 2 for more details of these components)
- initial experiments into all aspects of (re-)contextualization to validate the context model and to provide a baseline for future evaluation – some of these experiments have already started and are reported below.

The research plan will be frequently revisited and re-aligned with the activities in the rest of the project as well as with the requirements identified in collaboration with the two application pilots (WP9 and WP10) and the work in the architecture work package (WP8). The remainder of this section outlines how we see these three strands of research progressing over the next six months.

4.1 A Context Model for Contextualized Remembering

As a foundation for the other work and also closely interrelated with the work on contextualization methods it is planned to develop an adequate context model for the contextualization tasks in the ForgetIT project. This work has already started and the following important questions are currently under discussion:

- **Which dimensions of context are most relevant for contextualization in ForgetIT?** In the introductory section we already presented Time, Location, Topics, Entity Space and Document Space as an initial set of dimensions to be considered in ForgetIT. We have started from the well-established context dimensions time, location and topics, where time is considered with two sub-dimensions of creation time and content time. We have added the entity space and the document space as as further dimensions, because we think the entities and the documents and information object is interlinked with, are very important for its interpretation and , thus, should also be considered for contextualization. It is expected that there might be still some dimensions or sub-dimensions added for the organizational use case, where it is expected that e.g. related processes and organizational structures can also play an important role for the interpretation of information objects. This could, for example, be modelled as sub-dimensions of the topic dimension.

- **For each of the selected context dimensions, which information/features should be captured and preserved?** After identifying e.g. related documents or related entities, it has to be decided, which aspects of the context objects to capture as part of the context (e.g. which properties of the entities or the full document or just summaries). This process of selecting adequate features is expected to create varying results depending upon the type of information object and the use case under consideration. We expect, however to be able to identify some core time travel representations for the different dimensions (e.g. for entities) that can be adapted for the individual cases. It will also be considered here, that the context itself might again raise contextualization problems: Over time, it might become difficult or impossible to interpret the information that has been stored as context information. Thus, in selecting the way that the context is represented we also have to take into account its robustness with respect to evolution.

- **How to balance between capturing sufficient context and keeping context information concise?** Sufficient context information is required, in order to support future interpretation. In contrast, it is also important to keep the preserved content compact and to avoid to much redundancy.

4.2 Prototype Contextualization Components

As previously mentioned the majority of work planned within this work package over the next six months will focus on producing prototype contextualization components for use within the ForgetIT framework. It is envisaged that at least four components will be developed:

- A component for determining the context of textual documents. This is likely to be based around information extraction and disambiguation techniques as discussed in Section 3.1.1. This will be a generic domain independent component to provide a baseline implementation before use case specific techniques are developed later in the project.

- A component for determining the context of multimedia documents. This component will be based around current state-of-the-art techniques in image processing as detailed in Section 3.1.2.

- Integration with the archive to allow for storing the context along with the documents. The majority of this work will fall within WP5, the main focus in WP6 will be ensuring that the requirements for context aware preservation are accurately described and provided as input to WP5.

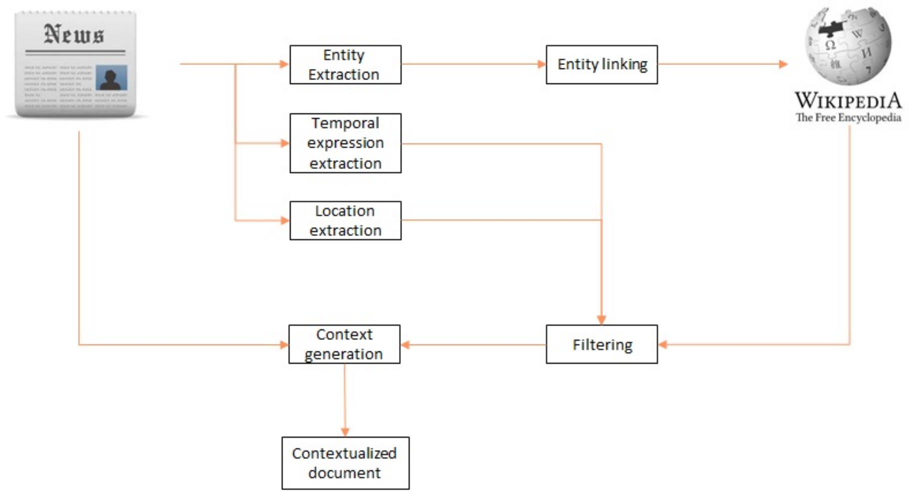

Figure 2: Block-diagram of the preliminary experiments in Re-Contextualization.

- A component to perform re-contextualization. An initial component to retrieve documents and their content from the archive will be developed and integrated into framework. This is likely, in the first instance, to be a simple implementation restores the items as archived. Further development of this component will depend heavily on the interfaces developed as part of the use cases. During the next research period requirements will be gathered to inform future development work.

It is worth noting at this stage that no mention is made of developing components for evolution-aware contextualization. While this is clearly an important part of the ForgetIT framework, it is difficult to envisage developing a component without having both a better understanding of the context model as well as components to determine context (i.e. you need to know what to evolve). The development of this component will, however, form a large part of the development work in this work package during the second year of the project.

4.3 (Re-)Contextualization Experiments

A further research line, which will be followed in the initial phase is re-contextualization, especially a-posteriori re-contextualization, i.e. the context is not stored at creation time, but it is re-constructed and interpretation time from external sources. These experiments are expected to give good insights into which information is required for a successful re-contextualization.

A preliminary experiment to investigate re-contextualizing a set of documents by exploiting an external knowledge base has already been started within the scope of this project. This experiment aimed to augment a given document - typically an older document - with external information in order to allow the reader to form a better understanding of the documents content. The datasets which are used in such investigation are the New York Times annotated corpus[5] (as document collection) and Wikipedia[6] (as knowledge base).

An high–level overview of the first experiments is depicted in Figure 2. References to named entities, temporal expressions and locations are extracted from the text of a given New York Times article. Then, the entities are linked to Wikipedia, i.e. they are mapped to their corresponding Wikipedia pages. Entities that do not have an associated Wikipedia page are simply ignored for the moment. Once the linking has been performed, the content of the Wikipedia pages representing the entities in the original article can be retrieved. This represents the knowledge pool from which the re-contextualizing information will be extracted. Clearly, not all the text of a Wikipedia page is related to the input article: we are interested only in those sections that are somehow related to the original context of the article. To this end, we use temporal expressions and locations extracted from the article to select the relevant sections. For the moment filtering units are sentences, i.e. Wikipedia pages are split into sentences, and a sentence is kept only if it contains either a temporal expressions or a location that is present in the original article. Finally, the extracted sentences or a summary of the extracted sentences will be combined with the original article in order to Re-Contextualize it.

The first experiments, as described above, focus on entities for the contextualization. As a next step it is planned to extend the approach to topics. This is expected to provide even more relevant re-contextualization information for an improved understanding. It is however, also expected that this form of contextualization will be more challenging due to the less clear identifiability of topics (as compared to entities). A further step in contextualization will include events as an additional feature of the context model. Given the entities mentioned inside a document, the basic intuition is that knowing the events in which the entities were involved might lead to a better understanding of the document. Finally, another line of extension of the approach is the inclusion of evolution aspects such as terminology evolution. Here methods will be shared with the work in evolution-aware contextualization (see above).

[5] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T19

[6] http://en.wikipedia.org/wiki/Main_Page

5 Conclusions

This deliverable has provided a concrete research plan for the next stage of the project. The research will focus on defining a formal model of context and implementing initial prototype components to implement the model. These components will initially focus on contextualization with later prototypes focusing more on the evolution of context and re-contextualization. This arrangement of work will allow us to provide basic components early within the project and incorporate feedback on them as well as information from the development of the use cases into future versions. These components will be based on a number of state-of-the-art techniques which are outlined in Section 3 and which should allow us to develop efficient and performant approaches which will benefit users of the ForgetIT framework.

References

[1] contextualize, v. In *OED Online*. Oxford University Press. http://www.oed.com/view/Entry/40215.

[2] Massimo Benerecetti, Paolo Bouquet, and Chiara Ghidini. Contextual Reasoning Distilled. *Philosophical Foundations of Artificial Intelligence. A special issue of the journal of Experimental and Theoretical AI (JETAI)*, 12(3):279–305, 2000.

[3] Doug Lenat. The Dimensions of Context Space. Technical report, Cycorp, 1998.

[4] Kalina Bontchevas, Niraj Aswani, Johanna Kieniewicz, and Michael Wallis. EnviLOD. http://gate.ac.uk/projects/envilod/.

[5] Hamish Cunningham, Valentin Tablan, Ian Roberts, Mark A. Greenwood, and Niraj Aswani. Information Extraction and Semantic Annotation for Multi-Paradigm Information Management. In Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 307–327. Springer Berlin Heidelberg, 2011.

[6] Mark A. Greenwood, Valentin Tablan, and Diana Maynard. GATE Mímir: Answering Questions Google Can't. In *Proceedings of the 10th International Semantic Web Conference (ISWC2011)*, October 2011.

[7] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M.A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (Version 6)*. The University of Sheffield, 2011.

[8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[9] Jaana Keklinen, Paavo Arvola, and Marko Junkkari. Contextualization. In *Encyclopedia of Database Systems*, pages 474–478. Springer US, 2009.

[10] Muhammad Ali Norozi, Paavo Arvola, and Arjen P. de Vries. Contextualization Using Hyperlinks and Internal Hierarchical Structure of Wikipedia Documents. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 734–743, New York, NY, USA, 2012. ACM.

[11] NEVAC. Thrust 1: Information Retrieval, Extraction, and Contextualization, 2009. http://www.geovista.psu.edu/resources/flyers/NEVAC_Thrust-1_GIR-Extraction-contextualization_final.pdf.

[12] Guoray Cai. Contextualization of Geospatial Database Semantics for Human–GIS Interaction. *Geoinformatica*, 11(2):217–237, June 2007.

[13] Gregor Leban and Marko Grobelnik. Displaying email-Related Contextual Information using Contextify. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.

[14] Mohamed Riadh Trad, Alexis Joly, and Nozha Boujemaa. Large scale visual-based event matching. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 53. ACM, 2011.

[15] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-Probe LSH: Efficient Indexing for High-Dimensional Similarity Search. In *Proceedings of the 33rd international conference on Very large data bases*, pages 950–961. VLDB Endowment, 2007.

[16] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[17] Annett Mitschick. Ontology-Based Indexing and Contextualization of Multimedia Documents for Personal Information Management Applications. *International Journal on Advances in Software*, 3(1 & 2), 2010.

[18] Nina Tahmasebi, Tereza Iofciu, Thomas Risse, Claudia Niedereé, and Wolf Siberski. Terminology Evolution in Web Archiving: Open Issues. In *Proceedings of the 8th IWAW*, 2008.

[19] Andrea Ernst-Gerlach and Norbert Fuhr. Generating Search Term Variants for Text Collections with Historic Spellings. In *Proceedings of ECIR*. Springer, 2006.

[20] Andrea Ernst-Gerlach and Norbert Fuhr. Retrieval in Text Collections with Historic Spelling using Linguistic and Spelling Variants. In *Proceedings of JCDL*, 2007.

[21] Klaus Berberich, Srikanta J. Bedathur, Mauro Sozio, and Gerhard Weikum. Bridging the Terminology Gap in Web Archive Search. In *Proceedings of WebDB'2009*, 2009.

[22] Nattiya Kanhabua and Kjetil Nørvåg. Exploiting Time-Based Synonyms in Searching Document Archives. In *Proc. 10th Annual Joint Conf. on Digital Libraries*, pages 79–88, 2010.

[23] Amal C. Kaluarachchi, Aparna S. Varde, Srikanta Bedathur, Gerhard Weikum, Jing Peng, and Anna Feldman. Incorporating Terminology Evolution for Query Translation in Text Retrieval with Association Rules. In *Proc. 19th Int. Conf. on Information and Knowledge Management*, pages 1789–1792. ACM, 2010.

[24] Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *Proc. 24th Int. Conf. on Computational Linguistics*, pages 2553–2568. ACL, 2012.

[25] *Magenta Book*, chapter Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems, 2 edition, July 2012.