D1.1 Data Acquisition and Integration Methods

Deliverable Co-ordinator: **Wim Peters**

Deliverable Co-ordinating Institution: **USFD**

Other Authors: Marta Sabou, Arno Scharl, Heinz-Peter Lang (Modul), Amel Fraisse, Patrick Paroubek (LIMSI), Gerhardt Wohlgenannt (WU)

Version	Date	Amended by	Changes
0.1	01-10-2013	Wim Peters	First model draft
0.2	03-11-2013	Wim Peters	Extension of model on the basis of other partners' input and first deliverable draft
0.3	12-11-2013	Wim Peters	Second draft
0.4	14-11-2013	Wim Peters	Final version

Executive Summary

This deliverable presents work performed I tasks 1.1, 1.2 and 1.4.

On the one hand it refers to data acquisition and integration software.

On the other hand, it provides a conceptual overview of uComp's activities, methods, techniques and data types in the form of a data model. This model also contains an evaluation framework.

Table of Contents

1. Introduction

This deliverable covers two aspects of work performed within tasks 1.1, 1.2 and 1.4.

1. A documented set of methods applied to particular resources. This software deliverable is the Extensible Web Retrieval Toolkit (eWRT), and is accessible and documented on http://www.weblyzard.com/ewrt/
2. The conceptual structure of uComp's methods, techniques, workflows and evalutation. This will form the content of this written deliverable.

2. The uComp domain

The uComp universe consists of a set of complex architectures and workflows. The project's activities performed by the four partners form an intricate network of individual and collaborative efforts. We argue that, in order to get a clear overview of all data acquisition and integration activities, modelling the overall structures is the best option.

The consequence is that this deliverable gives this overview of the conceptual structure of uComp. It provides a conceptual map with clear task descriptions, modules and outcomes, which will have a number of uses:

- It will provide the partners with a useful roadmap according to which to structure their efforts and collaboration.
- It will function as a blueprint of uComp's evaluation framework.

It is of course not possible to provide a fully specified model at this moment. The general idea of this deliverable is to provide an incremental conceptual overview of uComp, which will be further extended/specified in the course of the project.

3. The uComp model

The basic point of reference for this is the model as depicted in the figure below and formalized in https://gate.ac.uk/ns/ontologies/ucomp-data-model.owl
The white boxes represent concepts, and the coloured boxes list the instances that have been integrated into the model until now. Because of space constraints the picture cannot capture all relations in the model and the reader is referred to the owl version which can be browsed with ontology tools such as Protégé[1].
In order to embed the model in the wider context of Linked Open Data and standardized descriptions vocabularies for semiotic and linguistic/terminological descriptions, some elements from these vocabularies are being introduced. For instance, the information produced by uComp has been linked to the de facto foundational ontology standard of Dolce[2].

The model captures the produced data and the methods applied in uComp, on the basis of which various aspects such as data structure and evaluation requirements have been defined. It functions both as a conceptual map of uComp, and a conceptual framework within which evaluation criteria has been associated with data, methods and techniques.
Where possible, an incremental number of links to Linked Open Data vocabularies have been added.

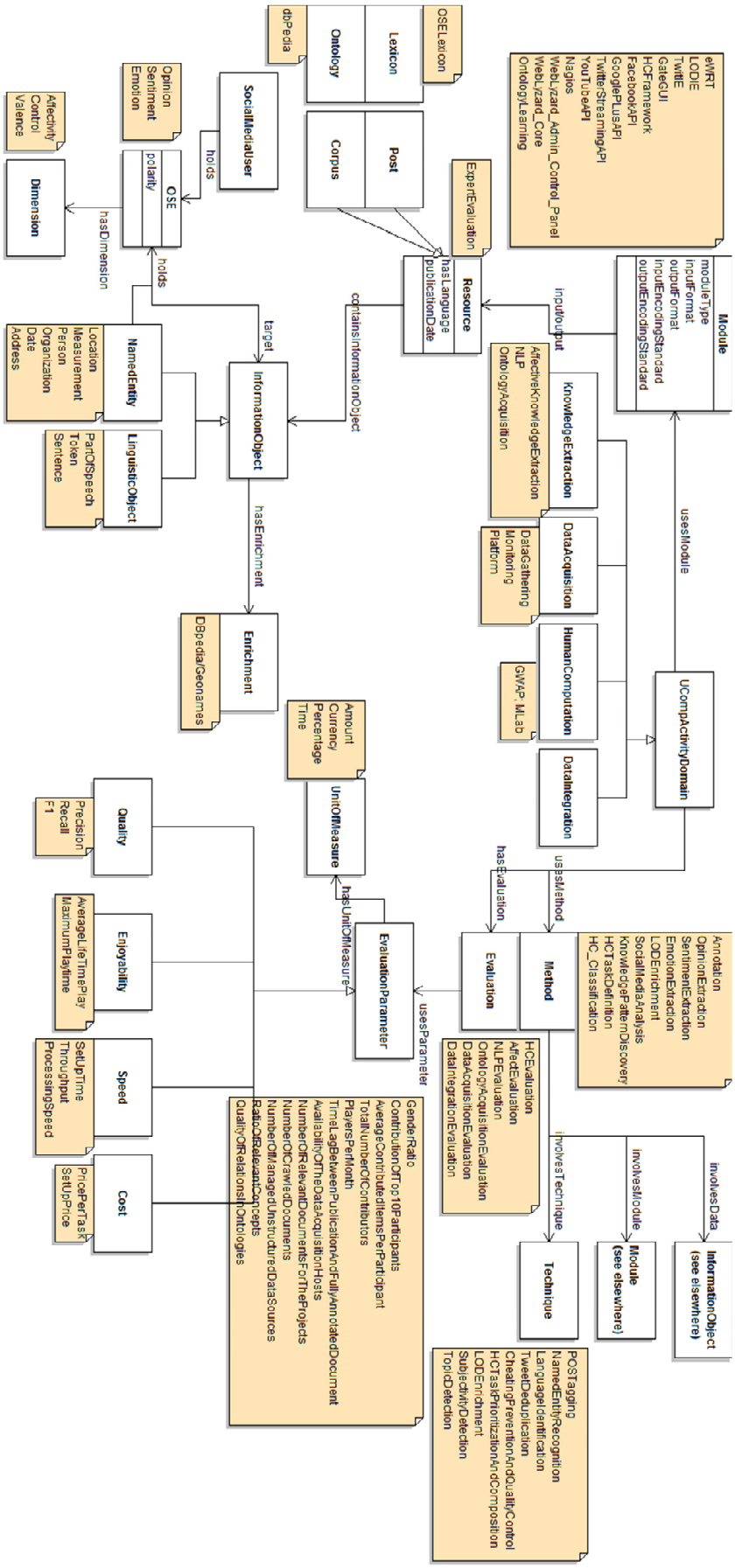
[1] http://protege.stanford.edu/

[2] http://ontologydesignpatterns.org/wiki/Ontology:DOLCE%2BDnS_Ultralite

It is important to note that this conceptual map provides an incremental specification of uComp's data and evaluation requirements. Although it has reached a stable and conceptually sufficient stage, it will be updated and refined as the project progresses and all open issues will have been addressed. This will have most impact on the description of uComp's methods and techniques, which have not yet been fully fleshed out.

As a general overview, the model's coverage of the informational and methodological structure is wide, but not exhaustive. It only covers the details of informational requirements to the extent that these are relevant for the overview it is intended to convey.
Moreover, it does not cover actual data, but only describes the data typology and associated evaluation strategies. For now, it is foreseen that the data structures produced by uComp will be encoded in XML. It will be decided later whether the encoding format will be extended to RDF. If this extension will take place, the extensibility of the data model will ensure that it will be able to accommodate the RDF description of all data types produced in uComp.

Figure 1. Conceptual overview of uComp



In the subsections below the uComp concepts, instances and their properties will be described. Concepts are always represented in bold, whereas properties and instances are mentioned in normal font.

3.1 Main concepts

Overall, this conceptual map is organized along six main dimensions, represented by the following concepts:

1. **UCompActivityDomain** (the main areas of work in uComp)
2. **Method** (methodologies and workflows applied within activity domains, used for structured and unstructured data acquisition and integration)
3. **Technique** (individual techniques applied for Methods)
4. **Module** (the method-specific computational tools that perform a.o. data gathering, analysis and provision. They produce/consume Resources within UCompActivityDomains and Methods)
5. **Resource** (the objects produced by uComp)
6. **InformationObject** (the produced metadata within uComp Modules and Resources)
7. **Evaluation** (aspects and coverage of the evaluation activities, including typology and **EvaluationParameters**)

These concepts form the operational scaffolding of uComp, and feature in the activities across the work packages under the individual or collective responsibility of the partners.

Figure 1 shows that the following relations exist between the main concepts:

Each **UCompActivityDomain** applies one or more **Methods**
Each **Method** involves one or more **Techniques**
Each **UCompActivityDomain** is associated with a particular **Evaluation** type
Each **Evaluation** applies one or more **EvaluationParameters**
Each **UCompActivityDomain** uses one or more **Modules**
Each **Module** produces zero or more **Resources**

3.2 Specific concepts

1. UCompActivityDomain represents a typology of activities performed by the partners.
The main subtypes are the following:

DataAquisition: Obtaining structured and unstructured documents and web objects
Instances:

A) **DataGathering**: the activity of acquiring structured and unstructured data according to a data collection plan
B) **Monitoring**: the activity of tracking the progress of data acquisition

DataIntegration: the packaging of information from different sources for uComp purposes
Instances:

A) Integration

HumanComputation: Activities/work flows that include human input in the production pipeline.
Instances:

A) **GWAP**: games with a purpose

B) **Mlap**: Mechanized Labour

KnowledgeExtraction: the acquisition of meaningful information.
Instances:
A) Affective knowledge extraction
B) NLP-based
C) Ontology acquisition

2. Method
Instances:
A) Annotation
B) Crawling
C) EmotionExtraction
D) SentimentExtraction
E) OpinionExtraction
F) HCTaskDefinition
G) SocialMediaAnalysis
H) KnowledgePatternDiscovery
I) HC_Classification

3. Technique
Instances:
A) POSTagging
B) NamedEntityRecognition
C) LanguageIdentification
D) TweetDeduplication
E) CheatingPreventionAndQualityControl
F) HCTaskPrioritizationAndComposition
G) LODEnrichment: enrichment of extracted named entities with DbPedia[3] and Geonames[4] identifiers. See also the **Enrichment** concept.
H) SubjectivityDetection: detects if a language sample expresses opinions, sentiments or emotions as opposed to purely objective text.
I) TopicDetection

4. Module
In order to maximize the flexibility of modelling, modules are associated with **UCompActivityDomain, Method and Technique.**

***Data properties*:**

- moduleType**:** whether created within uComp, external or a mix of internal and external
- inputFormat: an informal description of the input format.
- outputFormat: an informal description of the output format.
- inputEncodingStandard/outputEncodingStandard refer to any standard applied to the formulation of both input and output, for instance UTF-8, CSV, JSON

Instances:

A) eWRT
B) LODIE

[3] http://dbpedia.org

[4] http://www.geonames.org/

C) TwitIE
D) GateGUI
E) HCFramework
F) FacebookAPI
G) GooglePLusAPI
H) TwitterStreamingAPI
I) YouTubeAPI
J) Nagios
K) WebLyzard_Admin_Control_Panel
L) WebLyzard_Core
M) OntologyLearning

5. Resource: Data resources that are produced or consumed in uComp.

a) Lexicon: equivalent to http://purl.org/linguistics/gold/Lexicon
In its most general sense, the term is synonymous with vocabulary. A dictionary can be seen as a set of lexical entries. In linguistics, we don't normally speak of the vocabulary of a particular language; instead, we speak of the lexicon, the total store of words available to a speaker. Very commonly, the lexicon is not regarded merely as a long list of words. Rather, we conceive the lexicon as a set of lexical resources, including the morphemes of the languages, plus the processes available in the language for constructing words from those resources. Apart from the lexicon of a language as a whole, psycholinguists are interested in the mental lexicon, the words and lexical resources stored in an individual brain.

Instances:

A) FineGrainedOpinionsSentimentsAndEmotionsLexiconFor7Languages
A list of words. Each word has 2 attributes:
Polarity : will hold valence information (positive, negative)
Semantic Category : up to n (1 <= n <= 3) semantic categories taken from the 20 uComp OSE classes.
Example :
good, positive , [VALORIZATION, SATISFACTION, APPEASEMENT].
boring, negative, [BOREDOM].

b) Ontology: Equivalent to http://en.wikipedia.org/wiki/Ontology_(information_science)
In computer science and information science, a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts that can be used to reason about the entities within that domain, and may be used to describe the domain.

Instances:
A) DBpedia

c) Corpus: A collection of documents/texts.

d) Post: a social media submission, e.g. a tweet.

e) OSE_AnnotatedDocument: Language Samples annotated with uComp OSE annotations.

f) ExpertEvaluation

6. InformationObject
This concept originating from the IOLITE ontology[5] covers semantic data types that are produced by the uComp modules. These types are structured in the following way:

a) LinguisticObject: From IOLITE: **"**A linguistic object consisting of a string (independently of its physical realization).
Its topological unity can change according to its physical realization: as a written realization, its boundaries are blank spaces, as a spoken realization, sometimes is silence, sometimes not, and higher order features intervene.
Grammatical notions, such as noun, verb, adjective, etc., are roles defined by a grammar, and words (or larger linguistic objects) can play those roles in a given language. E.g., the word 'share' can play both 'verb' and 'noun' roles in contemporary English, while the word 'come' can only play the 'verb' role in English, and the 'adverb' or 'conjunction' roles in Italian (but if we consider a word as only realized by phonemes, i.e. if we consider the oral realizations of 'come', there is no common word 'come' in the two languages)."

In uComp we only consider the linguistic notions that form the output of GATE's[6] analysis of unstructured text.

Instances:

A) PartOfSpeech: One of a group of traditional classifications of words according to their functions in context, including noun, pronoun, verb, adjective, adverb, preposition, and conjunction.
B) Sentence: The largest syntactically independent unit of grammar.
C) Token: An orthographic unit in text.

b) OSE: Captures the various notions of affective knowledge.

Instances:

A) Opinion
B) Emotion
C) Sentiment

c) NamedEntity: refers to a elements from the list below that are mentioned by name in text.

Instances:

A) Person
B) Organization
C) Location
D) Date
E) Address
F) Measurement

[5] IOLite (http://www.loa-cnr.it/ontologies/IOLite.owl) is an ontology of information objects and realizations, and functions as a plugin to DOLCE-Ultralite (http://ontologydesignpatterns.org/wiki/Ontology:DOLCE%2BDnS_Ultralite)

[6] GATE is an an architecture for language engineering that has been developed by the University of Sheffield team since 2000. http://gate.ac.uk

d) Dimension: the aspects of Affect as detailed in deliverable 5.1: “Requirements of Affective Knowledge Extraction”.

Instances:

A) Affectivity
The affectivity dimension relates to the degree of the affectivity over the opinion, sentiment or emotion. According to this dimension, we distinguish between intellective, affective-intellective and affective expressions e.g., approval (intellective) versus joy (affective) or satisfaction (affective-intellective) versus happiness (affective).

B) Control
The control dimension relates to the degree of power over the affect, and helps to distinguish emotions initiated by the subject from those elicited by the environment e.g., contempt versus fear; this has also been called the strength, dominance, or confidence dimension in other models.

C) Valence
Valence dimension refers to how positive or negative the affect is; this is also referred to as subjective feeling of pleasantness or unpleasantness.

7. Evaluation: This concept is the top node for the definition of uComp’s evaluation framework.
The subconcepts denote various types of evaluation that are associated with instances of **UCompActivityDomain**.

Instances:

A) HCEvaluation
B) AffectEvaluation
C) NLPEvaluation
D) OntologyAcquisitionEvaluation
E) DataAcquisitionEvaluation
F) DataIntegrationEvaluation

Each evaluation instance makes use of **EvaluationParameters**.
For each activity/method/technology these parameters define concrete measures for qualitative and quantitative evaluation. A number of parameters have been subclassified under the headers of Quality, Cost, Enjoyability and Speed. Wherever possible parameters have been associated with **Evaluation** types, **Methods** and **Techniques** in the Owl model.

a) GeneralParameter
Instances:

A) GenderRatio
Percentage of women in player population.

B) ContributionOfTop10Participants
Percentage of tasks performed by the top 10 contributors (e.g., the top 10 scoring players in games or the 10 most active workers in a crowdsourcing project)

C) AverageContributedItemsPerParticipant
The average number of items/unit tasks (e.g., labels, rankings) performed by one participant

D) TotalNumberOfContributors

The total number of contributors in a HC system. This parameter is particularly interesting for GWAPs as a way to assess their success.

E) PlayersPerMonth
Number of players registering per month as a way to assess the success of advertisement.

F) TimeLagBetweenPublicationAndFullyAnnotatedDocument
Average time it takes from a new publication until our architecture acquires it and completes the annotation pipeline.

G) AvailabilityOfTheDataAcquisitionHosts
Up-time and availability for all hosts involved in the content acquisition pipeline throughout the whole project cycle.

H) NumberOfRelevantDocumentsForTheProjects
After applying the relevance and redundancy check, this number defines the relevant document for each source (e.g. news, blogs, Twitter)

I) NumberOfCrawledDocuments
Total number of documents from relevant project-sources grouped by type (e.g. news, blogs, Twitter)

J) NumberOfManagedUnstructuredDataSources
Total number of unstructured data-sources relevant for the project grouped by their type (e.g. news, blogs)

b) Quality parameter

Instances:

A) LexiconQuality
Size, linguistic coverage, consistence of annotations etc.

B) F1/Precision/Recall

C) QualityOfPrioritization

D) OSE_Annotation
Set of quality measures for OSE annotated corpus (quantitative objective measures, e.g. precision, recall, F-mesure for OSE annotation)

E) QualityOfRelationsInOntologies
Measures the quality of taxonomic and non-taxonomic relation detection in ontology learning. Suggests relation tyes for unlabelled relations

F) RatioOfRelevantConcepts
Measures the an aspect of quality of the ontology learning system. the higher the ratio of relevant concepts suggested by the ontology learning system, the better.

c) Cost parameter

Instances:

A) PricePerTask

The average amount of money spent for solving a problem instance (e.g., can be computed by dividing the total money paid to workers for a job to the number of problem instances that make up that job).

B) SetUpPrice
Captures a financial aspect of instances of HumanComputing subclasses (= the amount of money needed to design and set-up and HC application such as a Game or a job on MTurk)

d) Enjoyability parameter

Instances:

A) AverageLifeTimePlay

The overall amount of time the game is played by each player averaged across all people who have played it

B) MaximumPlayTime

The maximum amount of time the game is played by a given player

e) Speed parameter
Instances:

A) ProcessingSpeed
B) SetUpTime
The amount of time invested in building the HC application (including testing)
C) ThroughPut
The average number of problem instances solved, or input-output mappings performed, per human hour. (Some measure the amount of collected contributions per hour before these are aggregated into actual task solutions.)

In addition to the information conveyed on the picture, the data model further specifies the following:

- Each **EvaluationParameter** uses a **UnitOfMeasure** (amount, percentage etc.)
- Each **EvaluationParameter** applies to **Evaluation**, **Method** and **Technique**. This is not detailed in the picture due to spacing limitations, but is fully specified in the ontology.

8. Enrichment

The concepts that captures the linking of textual elements with elements from the LOD vocabulary

Instances:

A) Dbpedia
B) Geonames

3. Conclusion

Although the model is not yet complete, uComp's data acquisition/integration and evaluation strategies are clearly described through the formal structure of the ontology.

Ii is foreseen that in the course of the project, the various ontology elements will be revised, extended and complemented in order to arrive at a complete picture of uComp.

Acknowledgement: uComp receives the funding support of EPSRC EP/K017896/1, FWF 1097-N23, and ANR-12-CHRI-0003-03, in the framework of the CHIST-ERA ERA-NET