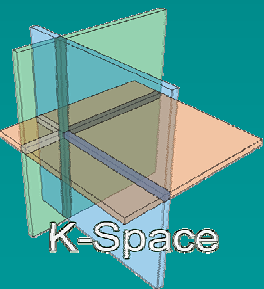


2nd Workshop on Challenges and Promise of the Semantic Web

# Semantic Technologies for Multimedia Analysis and Applications

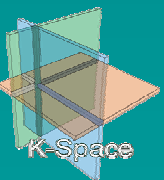
Thierry Declerck (DFKI GmbH)

With contributions by lecturers of SMSS 07,  
mainly Noel O'Connor and Alan Smeaton (DCU), and  
Alex Hauptmann (Carnegie Mellon)



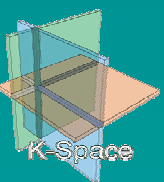
# A global Remark

- I am not talking about THE Semantic Web but rather about semantic webs, or even better about semantic technologies for multimedia
  - Till now no large scale Web applications of Semantic Web technologies. „Everyday“ web users do not tend to handle with Semantic Metadata, neither in generation of Web content nor in searching the Web (this being valid for both text and AV content)
- Great interest for semantic technologies in specific domains, like:
  - Bio-medicine (relation detection and extraction from text, with now an increasing interest on semantic annotation/processing of images).
  - Multimedia archives of press-archives
    - *Images etc are heavily annotated with text, using proprietary formats. There seems to be an added-value in providing the archives with Semantic Metadata.*
  - Etc.



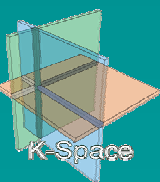
# The K-Space Project

- **K-Space -- Knowledge Space of Shared Technology and Integrative Research to Bridge the Semantic Gap** (in Multimedia Representation and Processing)
- A Network of Excellence in the 6<sup>th</sup> European R&D Framework (see <http://www.k-space.eu/>)
- K-Space is a network of research teams from academia and industry to conduct integrative research and dissemination activities in semantic inference for (semi-)automatic annotation and retrieval of multimedia content.
- K-Space is aiming at closing the semantic gap between low-level content descriptions that can be computed automatically by machines and the richness and subjectivity of semantics in high-level human interpretations of audiovisual media.

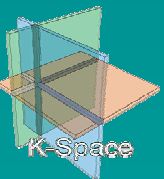
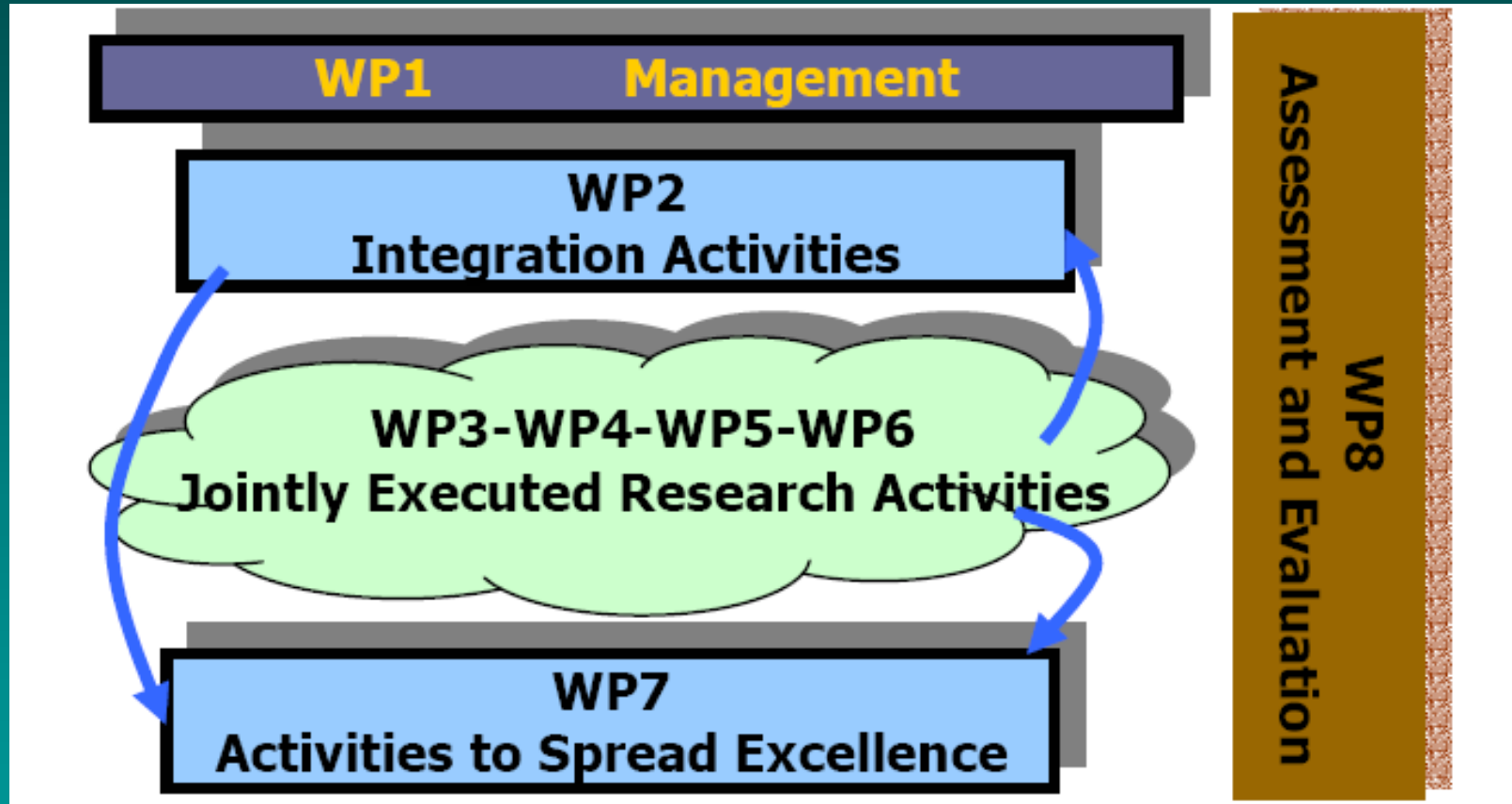


# Overview of Partners

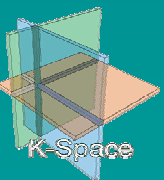
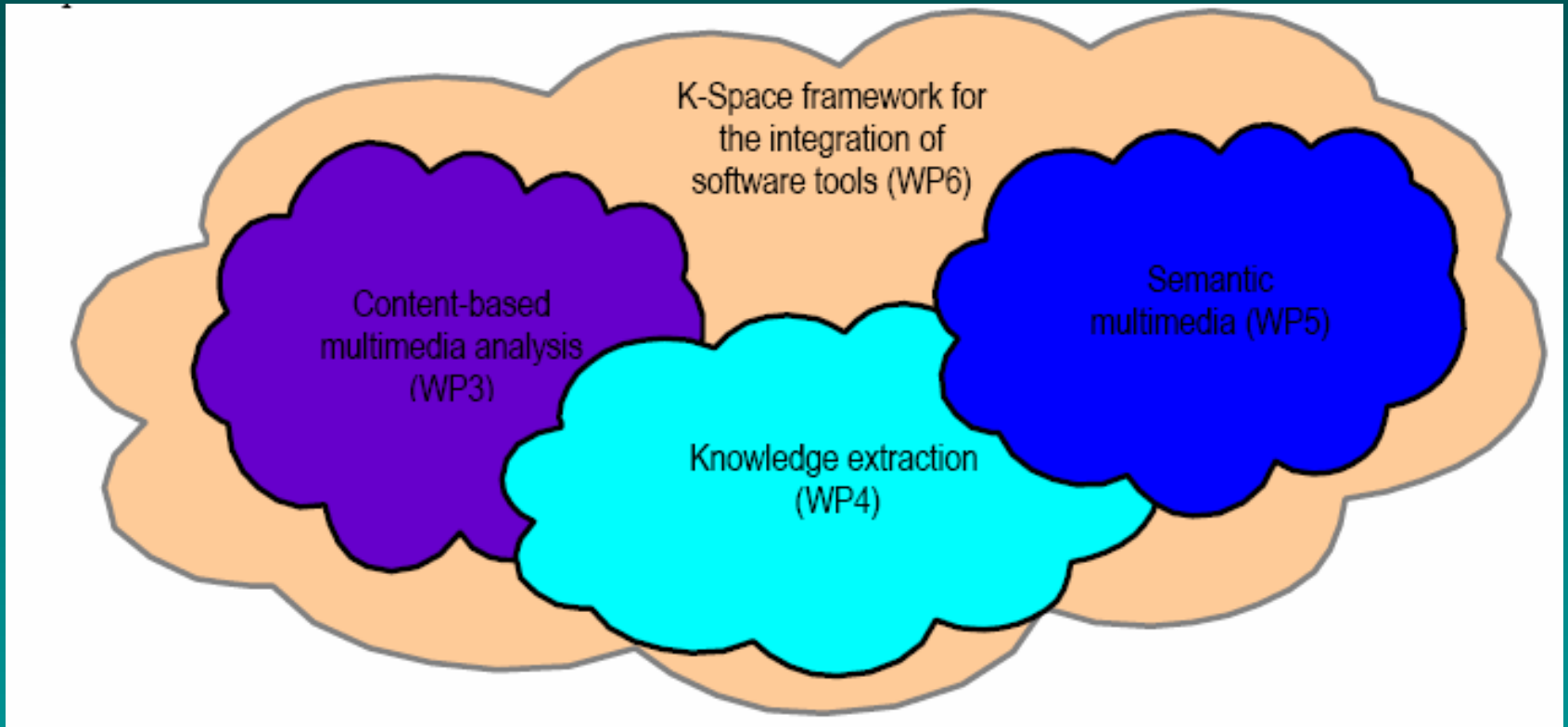
- Queen Mary, University of London, (QMUL)
- Koblenz University, (Uni Ko-Ld)
- Joanneum Research Forschungsgesellschaft mbH, (JRS)
- Informatics and Telematics Institute at the Centre for Research and Technology Hellas (ITI-CERTH)
- Dublin City University, (DCU)
- Centrum voor Wiskunde en Informatica, (CWI)
- Groupe des Ecoles des Télécommunications, (GET)
- Institut National de l'Audiovisuel, (INA)
- Institut Eurecom, (EURECOM)
- University of Glasgow, (GU)
- German Research Centre for Artificial Intelligence, (DFKI)
- Technische Universität Berlin, (TUB)
- Ecole Polytechnique Fédérale de Lausanne, (EPFL)
- University of Economics, Prague, (UEP)



# K-Space Framework (1)

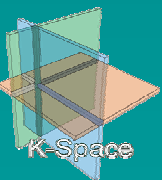


# Jointly Executed Research Activities



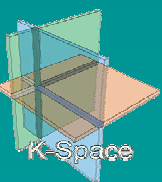
# Some Motivation (Noel O'Connor)

- Unprecedented growth in the amount and nature of multimedia content available
- Largely useless unless we can efficiently access relevant content
- Manual indexing of multimedia content for retrieval is time consuming & expensive
  - 8 hours for 1 hour of video!
- Content-based Information Retrieval (CBIR) Indexing performed automatically (ideally!)
  - Based on depicted 'real world' content - who, what, where, why, when, how, ...
- Clear need for machine computable techniques for extracting multimedia semantics
- Other reasons: Advanced interaction, improved compression, 'intelligent' content, ...



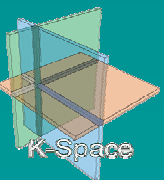
# Some Questions

- What is the main interest of „ordinary user“ searching the web for image/videos: primarily Semantics/Meanings, or satisfying perceptual, emotional or educational needs?
- Where is the semantics in Multimedia Analysis?
  - In the data (a panelist at SMSS 2007 in Glasgow).
    - *But then how to access/detect/extract the semantics? Via statistical mean only (in the broad sense), or data driven. Is this compatible with the aim of a „knowledge driven analysis of Multimedia“?*
  - Outside of the data, on the base of information/knowledge extracted from complementary data, like text, easier to map to domain knowledge (often encoded in the form of ontologies)
    - *But then how to relate this knowledge to the low-level results of multimedia analysis.*
  - An urgent need for a knowledge infrastructure to help in combining data-driven and knowledge-driven approaches, as well as bottom-up and top-down methods in the fields of Semantic Multimedia.



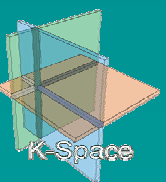
# Some Challenges

- There is also a need to identify precisely applications for which Semantic Multimedia is really proposing an added-value, and within those applications to precisely identify the contribution of the various elements playing a role in a Semantic Multimedia Infrastructure.
  - So for example pure image/videoretrieval tasks seem to be feasible without addition of knowledge/ontologies
- Problems in offering an thorough evaluation framework for Multimedia Semantics (see the huge effort deployed in TrecVid for evaluating the Summarization task in TrecVid 2007).
  - Also the hint that baseline systems seem to perform better than sophisticated systems ... (see SMSS 07 lecture by Alan Smeaton)



# More Questions

- Is MM Semantics „barely“ the association of concepts to images/video, or are there specific semantics to be extracted from low-level features (but then the challenge: how to extract it!)
  - Is the representational/reasoning power of semantic technologies to be applied to semantic metadata or also to MM content (fuzzy reasoning required here in any cases)
- What is exactly the content of an image, what should be annotated by semantic means? How valuable is the “subjectivity of semantics in high-level human interpretations of audiovisual media“?
  - See next slide on the human annotator (dis)agreement while annotating images.
  - See the 3 following slides: What should be annotated with the concept tennis?



# Annotator (dis)agreement



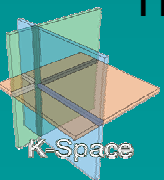
Possible concepts from automated visual Analysis => **Sky, Sea, Sand, Person**

Human annotator sometimes do not mention the sky (6 out of 20 annotators), but 5 persons have annotated the image with the keyword “sun”! (try also the query “beach, sea, person, sun” or similar in Flickr....)

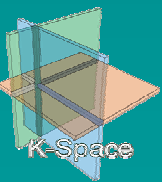
Human annotator gives very different details about the persons (friends, gender, pairs, etc.). Not really detectable by visual analysis.

So there is a need to define precisely what we mean by the “semantic gap”, in order to be able to evaluate performance of semantic extraction in multimedia processing.

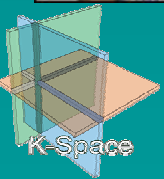
Picture: Courtesy of  
Thanos Athanasiadis (NTUA)



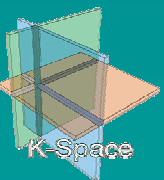
# Tennis concept adequacy??



# Tennis concept adequacy?

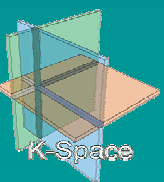


# Tennis concept (and subconcepts) adequate



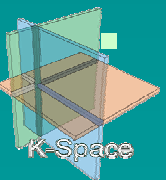
# Noel's definition of the Semantic Gap

- From signals to semantics:
  - Image Processing
    - *image in -> image out*
  - Image Analysis
    - *image in -> measurements out*
  - Image Understanding:
    - *image in -> high-level description out*
    - *[Young et al, Image Processing Fundamentals, CRC Press LLC, 1997]*
- Problem:
  - Difference between what we can measure from a visual signal and what this means
  - The “Semantic Gap”
    - *[Smeulders et al, Content-based image retrieval at the end of the early years. IEEE transactions PAMI, 22 -12:1349 - 1380, 2000.]*



# Meanings in Multimedia?

- Nearly the whole world/universe can be represented in images
  - Meaning: The background knowledge for the semantic annotation (encoded in taxonomies, ontologies) is huge, potentially infinite
  - Need to restrict the conceptual „space“ to the domain, application, task under consideration
- Concepts attached to raw image/video material, or to compressed image/video?
- Concepts attached to regions or to the whole image image? Is there a compositional semantics for computing the semantics of the whole images out of the semantics of the regions of the image?
  - Compositionality seems to be playing a role rather in video analysis (over the temporal dimension)
- Are there main elements/regions/frames in term of saliency of semantics? (this kind of structure, so-called „dependency structure“ is central to textual analysis, see next slide)

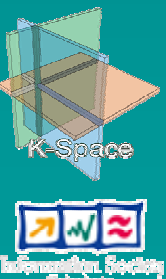


# Example of Dependency Structure in Text

- Roger Federer holding the Trophy
  - [<sub>NP</sub> Roger Federer *subj*] [<sub>VG</sub> holding *predicate*] [<sub>NP</sub> the Trophy *direct object*]
  - The dependency holds for the the predicate „holding“, and the subject stays in the holding relation with the direct object. So here we know that we can see Federer, but not during a playing event.

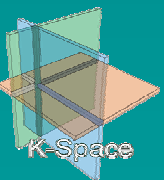
# Meanings in Multimedia?

- Two important semantic descriptors not (always) in the data
  - Time information seems to be intrinsically outside of the image material (can analysis tools recognized dated images on the base of low-level features?)
  - Precise location, like (name of) countries, cities are also mostly gained from information available outside of the image.
    - *Not all relevant spatio temporal information/knowledge to be extracted from the image, using only image/video analysis tools (but with the help of text included in the image, or speech/text associated with the image/video)*
  - Need for an integration of information/knowledge originated from various sources, media and modalities.



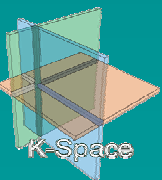
# Meanings in Multimedia?

- Does the image/video has the difference between „intension“ and „extension“?
  - The say „an image is worth thousand words“ is valid for the perception of humans. But the inverse „a word can be represented/expressed by thousand, and more, images, is certainly also true. This is due to the intension/extension distinction.
    - *Intension = intrinsic meaning“ of a word,*
    - *Extension = all the objects in the world denoted by this object.*
- Is the distinction type/token also valid in Multimedia?



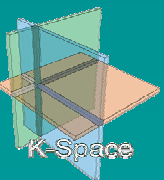
# A good summary by Noel O'Connoer

- Extracting semantics from a 2D array of pixels is hard!
  - How can we make life easier?
  - Design capture devices that make the problem easier ...
  - Augment visual sensor with other modalities
  - Examples:
    - *Multiple cameras (e.g. stereo)*
    - *Beyond the visible spectrum (e.g. infrared)*
    - *Non-content sensors: time, date, location, movement, ...*
- But (my comments):
  - Multimodal approaches have their own problems as well, for example, the non aligned segmentation of the various media/modalities
    - *Shot segmentation in AV signal is very often not aligned with the meaningful segmentation of the complementary speech. Speech can thus not be (optimally) used for adding semantics to the shot.*
  - Better use text present in the image in this cross-media approach, but OCR still is a bottleneck (detection of text regions in images seem to work better)



# Maybe a good and ambitious scenario for Semantics in Multimedia (Again by Noel (SMSS 2007))

- Consider providing access to movie content
  - Shots still exist, but no nice (easy) regular repetitive “news story” structure
  - Film making is an extremely creative & artistic medium!
- Define movie events
  - A temporal segment that viewers recognise easily and remember as a semantic unit
  - An event may be a scene (or not)
- 3 important types of events
  - Dialogue, Exciting and Musical.
  - Account for > 90% of a film, whilst being intuitive for a user to understand
  - Studies indicate that events more intuitive than shots or scenes in a variety of retrieval applications



# A Proposal for a Bridging Semantic Infrastructure:

This part of the talk is to be found in an external set of slides, which was designed and presented by Paul Buitelaar for an invited talk given to the X-Media Project

