From Intelligent Content to Actionable Knowledge:

Research Directions and Opportunities under Framework Programme 7

Roberto Cencioni & Stefano Bertolo
European Commission, Information Society and Media
infso-e2@ec.europa.eu

Abstract

Since many human activities depend on the creation, use and transmission of symbolic information, advances in our ability to produce, organise, distribute and exploit such information (semi)automatically can be expected to have great impact on society and economy. For this reason, this objective has been proposed as one of the main activities under Framework Programme 7 (FP7), the next cycle of EU research and technology development activities to run through 2007-2013. This paper gives a broad overview of the place of content and knowledge research within FP7 and discusses several lines of research that have been identified as particularly important and promising.

Motivation

Even an extremely superficial look at the history of technology shows that as soon as a human activity becomes of economic or societal significance it becomes important to make it more efficient. Depending on the circumstances this might mean either allowing for greater output to result from the same amount of activity or by allowing more people to participate in the activity by reducing its physical or mental demands. This sounds almost too trivial to mention. Consider however that in developed countries, and in the EU in particular, a large and growing number of people is daily engaged in the process of creating, storing, distributing, accessing and reasoning about what one could in the most general terms describe as content (documents, pictures, videos, music …) and knowledge (descriptions of reality that can be used to summarise past experiences and/or draw conclusions that go beyond them). But while we have efficient tools to, say, lift crates or cut metal to a desired shape and adequate infrastructures to extract and transport energy, our existing tools for managing content and knowledge (semi)automatically are, by comparison, still extremely primitive, with the possible exception of textual indexing and search.

When you come back from a trip and your digital camera disgorges on your hard disk dozens of sequentially numbered pictures you still have to manually inspect them to make better sense of their content. What is a minor inconvenience to an amateur photographer becomes a major bottleneck in the work of a professional reporter or newsroom editor. Similarly, while doctors keep records of their patients' background and conditions, only highly trained medical professionals are able to compare them meaningfully to draw appropriate diagnostic conclusions. If some of these comparisons could be carried out automatically and brought to the attention of the physician when appropriate he could devote more of his time to other aspects of his practice.

These two simple examples show how different our activities would be if we could create a global infrastructure to do for human knowledge and digital content what early 20th century industrialization did for electricity, allowing it to be produced, transmitted and used reliably, efficiently and according to standards that could be relied upon to build applications and tools. This vision is the motivation behind FP7 plans for content and knowledge research and development activities.

A Brief Overview of Framework Programme 7

Multi-year programmes called Framework Programmes have since 1984 been the European Union's main instrument for funding research and technology development. As the current Framework Programme 6 (2002-2006) draws to an end, the European institutions have been at work defining Framework Programme 7 with the Commission's proposals going through the co-decision procedure for approval and adoption by the European Parliament and Council. At the date of writing the co-decision procedure is still in process and is expected to come to a conclusion at the end of 2006, at which time FP7 will be formally launched and its *work programme* published and organised into a series of *objectives* that will form the object of calls for proposals.

According to the most recent figures, in FP7 around € 9000M will be devoted to multi-party, multi-nation research activities in information and communication technologies (ICT). It is anticipated that nearly € 250M will be devoted to content and knowledge themes in the period 2007-2008, with several calls for proposals expected to be published in 2007 covering topics such as digital libraries, networked media, multimedia content, semantics and knowledge management. The topics selected are the result of articulating the vision described in the previous section against the background of the Commission's i2010 policy initiative, the recommendations received by various bodies of experts such as the IST Advisory Group (ISTAG), extensive consultations with researchers and technology developers from all over the European Union, and constant monitoring of global technology trends to identify gaps and opportunities.

FP7 Research Directions in Content and Knowledge

While at the time of writing they have not yet been finalised in a formal work programme, it is safe to anticipate that FP7 directions in content and knowledge can be seen as an evolution of FP6 efforts with continuity assured for hard problems that are still considered open, acceleration in areas in which external developments and dynamics invite more decisive action, and new lines of activity for topics whose importance and promise could not be fully appreciated just a couple of years ago. These are not necessarily in alternative to one another and indeed novel lines of activity often suggest new lines of attack for long standing problems as is the case for social network analysis and knowledge management.

Acceleration and novelty are due to the enormous changes in the production and consumption of content and knowledge that have taken place from just a few years ago. While past Framework Programme efforts have often concentrated on formal knowledge extracted from textual and multimedia sources, it has become evident that these are not going to be the only forms in which content and knowledge will exist and be consumed.

The first trend that can be readily observed is the explosion in the availability of networked multimedia content and the fact that much content is produced and remixed by non-professionals and accessed/consumed on devices of great variability in terms of networking capability, display resolution and controls. This brings about problems of scale, usability and democratisation of the tool chain that need to be addressed.

The second trend, which we are already witnessing in the most advanced scientific laboratories but can soon be expected to spread to many other environments is the proliferation of data that has been produced by devices as opposed to humans. Such data are in need of interpretation and integration and, for reasons discussed below, they create novel demands on knowledge representation and reasoning.

A further clear trend that has guided the formulation of the FP7 content and knowledge objectives is the speed at which distributed (e.g. peer to peer) and socially enhanced content management applications have established themselves as successful solutions, throwing into sharper relief than it was possible in the past issues such as trust and provenance, personalisation and contextualisation.

Before delving into a number of research themes that have been identified as central to FP7, a few methodological remarks that apply to all of them are in order. Two of the corner stones of the scientific method are accuracy of measurement and replicability. In the context of upcoming FP7 activities accuracy of measurement means that proposed research and development activities should come with clear plans as to what aspects of the system's performance will be measured and how the measurements will provide evidence that the benefits of the system outweigh its costs. This is to cover not only algorithmic performance but also issues of usability. Replicability means that developers are expected to demonstrate their system's performance in conditions that can in principle be replicated elsewhere by other groups so as to invite comparison and provide the means for progress tracking and cost benefit analyses. Additionally, those conditions will need to be realistic, i.e. address data collections and user populations of the same order of magnitude expected in the eventual deployment of the technology developed. This means that in FP7 projects, securing access to realistic data collections and relentlessly benchmarking and testing against those collections can be assumed to be as important as working on ideas that are scientifically or technologically novel.

We are now ready to give a short description of specific research lines, their motivation and their expected outcome. They are listed in no particular order.

Authoring Environments

As mentioned above, one of the obvious trends of the last couple of years is the enormous growth of user produced multimedia content that is presumably the effect of both the increased availability of digital cameras and the ease with which the materials so produced can be uploaded and shared on the Internet.

Between creating/capturing and sharing there are however various forms of planning and editing that need to be better supported. These go from various forms of image correction to enhance the appearance of people or objects, to novel forms of storyboarding, to the remixing of content from previously existing materials, to the creation of animated models from real life images. Recognizing objects of interest in digital content and making them available for manipulation in an editing environment according to their specific type (much as programming constructs can be recognised and manipulated in an integrated development environment) would certainly benefit professional creators of content.

Activities under this theme however will also be required to be mindful of two social dimensions. The first one is the democratisation of the tool chain: bringing these advanced content editing functionalities within the technical reach of casual users can be expected to have a positive effect on the widespread creation of content of better quality. Once this content is created, it is important that it be shared and searched effectively both locally by the user as a form of personal data management and in collective environments, whether organisations or communities. The second dimension is the creation of open standards and tools for (semi)automatic metadata annotation in support of more sophisticated forms of both symbolic and similarity based multimedia search.

Workflow Environments

More directly targeted in support of organisations and professional creators of content will be activities to develop collaborative workflow environments in support the entire lifecycle of media and enterprise content. These should go from the planned acquisition of raw materials (especially from legacy collections) to the versioning, packaging and repurposing of complex products allowing for content and annotations produced on a given application to be saved and stored according to open standards to be reused by other applications in the workflow.

The technology developed should be able to address the needs of extremely large multimedia archives, appropriately addressing physical storage and indexing schemes. This is considered essential to ensure a transition of the technology from the research to the actual deployment stage in a commercial environment or a digital library. One of the dimensions of repurposing will naturally be the (semi)automatic preparation of content for various target audiences with widely varying bandwidth and display capabilities, and different cultural and linguistic requirements. Under this dimension, attention will be given to efficient techniques for multimedia summarisation and encoding schemes based on the properties of the target device and the psychology of human attention and perception. This will allow for salient features of multimedia segments to be identified and for the non-salient features to be compressed with little or no loss in the overall experience at the point of consumption.

Personalisation and Consumption of Content

The success of the two previous research lines produces a scenario in which content objects will come into existence and be distributed with a considerable amount of actionable knowledge about themselves (what they represent, when and where they have been captured/created, …). This creates two significant opportunities at the point of consumption that will be explored in FP7.

The first one is to use the knowledge embedded into the object in interaction with knowledge of the user and emergent ambient intelligence, to support the most effective multimodal experience. In this scenario the user would select a content object and the object would determine how to best display itself and expose its controls, potentially co-opting various forms of hardware detected in its environment based on its understanding of the user (goals in the interaction, language and cultural preferences, …) and his current circumstances (location, time, …).

The second opportunity is to allow user interactions to flow back into the object and add to its intelligence in a sort of ecological validation. The object will be able to unobtrusively detect and record if and how it is being interacted with (looked at, clicked at, resized, skipped …) or even what emotions it has elicited. This information, when made available to the content creators, will allow them to edit the object for improved effect, just as web designers today improve the usability of their sites after analysing click-through logs or eye-tracking heatmaps.

Both opportunities clearly depend on accessing or collecting a non-trivial amount of personal information on content consumers and their behaviour. Projects addressing the delivery of smart content objects will thus need to appropriately address consumer privacy issues, as discussed in the next section.

Semantic Foundations

Formal knowledge representation remains very much central to the FP7 research efforts but its emphasis will be focused towards demands that have been identified from an analysis of the types and scale of the data that it is expected to integrate. The massive amounts of structured information coming on-line as the valuable output of government bodies (e.g. census and cadastral data), research laboratories (a trend most obvious in the life sciences) or any other source, will make it ever more important to build and manage the semantic 'glue' that will allow reasoning over data stored in different databases, whose semantics are not directly aligned. This is indeed one of the central insights of the *Semantic Web* vision. This requires methods for aligning ontologies and support reasoning over distributed knowledge sources.

These large amounts of structured data moreover offer an obvious opportunity for theory induction, i.e. the ability to reach general conclusions from the analysis of a large number of individual observations. The ultimate goal for semantic integration and theory induction is to produce programs that routinely outperform trained professionals in their ability to reach important conclusions and produce insightful and novel hypotheses from the data at hand. The very heterogeneity of those sources however will mean that noise and inconsistencies will be inevitable: novel techniques in reasoning and machine learning will be needed to overcome them likely leading to an integration of symbolic and probabilistic knowledge representation.

Existing knowledge representation formalisms will also need to be extended to be able to represent and reason about objects that exist and change in time, and processes that unfold and branch over time. A second source of massive amounts of data that is just behind the horizon and for which appropriate knowledge integration solutions do not yet exist are sensors and physical objects endowed with various form of radio identifiers and locators. In this scenario too we will face massive streams of structured data that can be queried for the occurrence of certain events and from which much can be induced. Developing the technology for doing this efficiently and on a massive scale will be an important objective in FP7.

Knowledge Management Systems

The previous section addresses the knowledge representation problems that need to be solved in environments where data, while potentially noisy, are highly structured and typically interpretable against the background of mature sciences or well defined systems. Most organisations however operate in environments where those characteristics do not apply and where knowledge is typically stored in textual or multimedia documents of arbitrary format and content. These organisations need technologies capable of extracting *actionable meaning* from unstructured or poorly structured information as well as from social interaction patterns, and of making it available for activities ranging from information search through conceptual mapping to decision making.

The goal is to progressively integrate the organisation's knowledge assets into its formal processes and to be able to expose both to third parties in the dynamic creation of virtual organisations as required by common business objectives. The associated security concerns will be addresses by the definition, verification and automatic implementation of formal policies designed to regulate access to data sources or other kinds of organisational resources. Formal policies are indeed a prime example of the need for extending existing knowledge representation formalisms to include rules and temporal reasoning.

Progress in such advanced knowledge management methods and techniques will be tested in real-life settings with particular attention to the integration of legacy systems and to issues related to usability, scalability, flexibility and effectiveness. New approaches and technologies will be embedded within systems using computer-tractable knowledge in support of dynamic data and application integration, automation of organisational processes, automated diagnosis and problem solving in a variety of knowledge-intensive domains.

Socio-economic Studies

The significant increase of user produced content available on the Internet has been one of the most noticeable trends of the past few months but it is fair to say that it has been rather serendipitous with even the largest commercial players reacting to it rather than breaking new ground. The time seems ripe to undertake a systematic study of what economic and social factors act as catalyst or inhibitors to the production and distribution of user content and what information and communication technologies would be needed to enhance the effects of the catalysts and reduce the effects of the inhibitors.

For example, as seen in the previous sections, enormous benefits could be reaped from systematically collecting user feedback on the consumption of content but at the same time users cannot legitimately be expected to surrender this information indiscriminately. This line of research will thus encourage work on privacy preserving data mining algorithms for collecting this feedback from an aggregate analysis of social and user device interactions.

Similarly, users may decide to produce more content, or more content of a certain valuable type, if they could rely on certain forms of reward (not necessarily financial) or feel that their repurposing or remixing of pre-existing content is fully legitimate. Determining the appropriate mixture of incentives will require empirical studies in social psychology and economics.

Finally, this line of activities will foster community building to encourage multi-disciplinary approaches and a more effective dialogue between suppliers and users of technology, for a faster and broader uptake of research results.

Conclusion

All of the research directions described above can be seen as related strands of a simple goal: (a) identifying human activities where the constraining factor is the availability of human intelligence and (b) implementing technologies capable of amplifying it or replacing it when appropriate very much in the same way as physical machines amplify or replace human strength or dexterity. If these goals are met people will be able to concentrate on activities of ever greater depth and creativity, and considerable productivity gains will materialise in those information rich domains from which innovation has traditionally come, hopefully triggering a virtuous circle of global proportions.