A Brief Note on Knowledge Technologies

by Hans-Georg Stork, Luxembourg

Issues and challenges

Knowledge is a prerequisite for acting purposefully in a given environment or domain (decision making, planning, collaborating, etc.). Knowledge is always about something: objects, processes, phenomena, etc.

In order to be amenable to ***"automated* (i.e. computer-implementable) *solutions"*** knowledge must be formally represented. Computer-based **representations of knowledge** about objects and processes (digital or not) capture, to a certain extent, the *semantics* of these objects and processes (including e.g. people, papers, articles, reports, books, recipes, databases, still and moving images, graphs, product and service descriptions, etc.).

Adding (automatically or interactively) **explicit semantics to** (static) **content, services and processes** (and thereby producing knowledge representations) is one of the key functions of *Knowledge Technology* tools. They (help to) generate and record the ***"meta-knowledge"*** (about digital content of all sorts) that makes all other forms (such as scientific and scholarly papers) of knowledge (and often nonsense) more accessible and usable.

Content-/service-providers and end-users need to be made aware of the benefits of adding explicit semantics to content. This is largely a ***"critical mass"*** or "chicken and egg" problem: adding semantics to content (and services) does not pay off if no tools are available to make good use of it while developing tools does not pay off if there is little semantically-enriched content to work on.

Open source development of suitable software geared in particular towards very large scale open distributed systems for knowledge management and use, on platforms such as the World Wide Web, may be a viable ("piecemeal engineering") approach to achieving this critical mass. These developments could address simple applications, of interest to as many people as possible, such as personal information management (PIM) or e-publishing (scientific and scholarly preprints), demonstrating the value of adding explicit semantics to content.

"**Acting upon semantically-enriched content** (including service descriptions)" refers to a second important class of functions to be provided by *Knowledge Technology* tools and artefacts.

In order to make the most of semantically-enriched content in open distributed systems agents must "understand" each other, they must be ***interoperable***. This must be guaranteed at all levels: syntactic, semantic and pragmatic. Solutions to interoperability problems require specific models and tools (e.g. for **mappings** between representation formalisms/**standards**) and may also need specific organisational underpinning.

Interoperability is indeed only one (yet important) aspect of the *quality* of knowledge representations and tools. Other aspects include *scalability* and *usability*. To ensure the overall quality of such artefacts a coherent and well organised process of test, evaluation, assessment and certification is desirable.

Strong business cases for the cost-effective use of *Knowledge Technologies* in corporate and/or commercial environments must be further elaborated. Specific examples include

- workflow and collaboration support,
- proactive portals for community building,
- retrieval, filtering, profiling and recommender systems, as well as
- document change and innovation management.

Trustability, privacy and other **social issues** should be taken into account at both the design and operation stages of knowledge-based systems. Target areas such as "customer relationship management" (CRM) or "e-government" are particularly sensitive in this regard.

Knowledge Technologies draw on various Computer Science sub-disciplines such as formal modelling, logics and languages, information retrieval, (multimedia) databases, image analysis, cognitive vision, etc., but also on "trans-disciplines" such as Cognitive Science. Therefore R&D projects addressing *Knowledge Technologies* will necessarily be **multi-disciplinary**.

R&D areas

R&D areas can be categorized broadly as pertaining to **(i) adding explicit semantics to content, services and processes** or, **(ii) acting upon semantic descriptions** (cf. above).

(i) Adding explicit semantics to content, services and processes

As explained in the previous section, knowledge about content, services and processes is made explicit through formal *descriptions* of such entities. These descriptions are usually referred to as *metadata*. However, while formal description is necessary for agents to act upon, it is not sufficient. Descriptive terms are "understandable" (or "meaningful") only if their meaning has been defined somehow, somewhere. This is usually done through *ontologies* which provide meaning that can be operated upon. They embed terms in contexts (of other terms) and/or stipulate rules a given term (or set of given terms) must obey.

Metadata and *ontologies* demarcate broad and inter-related research areas. Pertinent problems and solutions depend largely on content types and usage environments. Problems include:

Metadata and indexing

- metadata extraction / capture
- semantic annotation
- domain-, context-, user- and task-oriented indexing
- semantic indexing of **multimedia content**

Ontologies

- ontology construction (including ontologies for multimedia objects) and management ("knowledge lifecycle" support)
- ontology learning

There are various classes of technologies and/or approaches likely to provide **generic** solutions, for instance:

- data/text mining for "knowledge discovery" in data bases or large text repositories
- concept detection and fact extraction
- machine learning (e.g. for automatic classification)
- semantic analysis of audiovisual content (segmentation, object extraction, etc.)
- speech, face, gesture and emotion recognition (e.g. through natural language analysis and cognitive vision)

While it would be preferable to automate completely processes of **"knowledge acquisition"** in or for knowledge-based systems, some **elicitation** component (e.g. for capturing knowledge through targeted man-machine interaction) will most likely always be required. **Adaptive** and **context-sensitive** techniques are needed. *Corporate knowledge* systems for instance should provide *enjoyable* means of interactive knowledge capture on the fly, directly from workflow processes.

In large **distributed systems** (such as Internet-, intranet- or extranet-based webs) knowledge creation and management pose particular problems and challenges. Specialised **services** (see also subsection (ii), below) could be offered to support both knowledge acquisition and knowledge elicitation (e.g. for semantic annotation of content). *Peer-to-peer* (P2P) networks (generalizing the by now classical client-server configurations) would lend themselves to implementing methods allowing semantics to **"emerge"** from node-to-node interaction.

Given the sheer amount of content in global distributed systems solutions to some of the above problems (e.g. semantic annotation, multimedia content analysis, etc.) may require powerful computing resources as provided for instance by *Grid* computing technologies. Grids may in fact contribute all kinds of compute-intensive services to be offered through webs (see below).

(ii) Acting upon semantic descriptions

Semantic content (service, process, ...) description (based on suitable ontologies) enables *software agents* in distributed systems to co-operate and to perform complex transactions and other operations (such as searching, filtering and integrating information), on their users' behalf and without extensive user intervention. Semantic descriptions relieve implementers of the burden of "hard-coding" semantics in agents and thus contribute to achieving interoperability. They can also help *human agents* ("users") to make sense of and interact with content, services and processes in distributed systems.

These general comments imply a wide range of challenging and **generic** R&D topics.

Some of these topics can be subsumed under the headings **"*Semantics- (or knowledge-)based services*"** and **"*service semantics*"**. Services are of particular relevance in distributed systems such as the World Wide Web. Indeed, the term **"*Semantic Web*"** usually refers to the formal framework (in terms of models and languages) needed to provide agents with semantic (i.e. ontology-based) descriptions of all sorts

of web-addressable entities (including services), allowing for instance context-sensitive *service discovery*, *mediation* and *composition*. These agents would also build on *reasoning/inferencing* capabilities that ontologies make possible. However, *scalability* of ontologies, ontological reasoning and ontology (change) management remain serious problems.

The **"interfacing with knowledge"** aspects (i.e. making content, services, etc., accessible and intelligible to people) are equally intriguing. Subjects falling into that category and allowing for largely generic R&D include:

- semantics based navigation and browsing
- semantic search engines with domain-, context-, user- and task-sensitive query construction support
- knowledge-(viz. semantics-)based dialogue management
- semantic Web portals and collaboration support
- user profiling, personalisation, customization (e.g. through particular "views on knowledge")
- visualizing knowledge
- device dependent interfacing.

Research has been conducted on many of these and similar subjects for quite some time (with or without the "knowledge" or "semantics" qualifier); yet they could benefit from new knowledge-based approaches. And they become ever more important and more challenging as the global (wired and wireless) networks increase their reach both in terms of capacity and physical access modes, evolving into an *"ubiquitous permeable web"*, necessitating **"semantics for everything"**.