D 1.4.2 Success Stories and Best Practices

Coordinator: Luigi Lancieri (FT)

Diana Maynard (USFD), Fabien Gandon (INRIA)

Abstract.
EU-IST Network of Excellence (NoE) IST-2004-507482 KWEB
Deliverable 1.4.2 (WP1.4)
In this deliverable some best practices and illustrative examples of success stories in semantic web applications are analyzed. In particular three main activities have been synthesized: (i) the definition of "practices" in the semantic web domain has been unveiled, analyzing a multiple choice questionnaire which is intended to evaluate the level of consensus among practitioners and researchers; (ii) the identification of the most important semantic web best practices, surveyed through the study of present and past initiatives related to W3C activities; (iii) the identification of some practices in concrete case studies.

Document Identifier:	KWEB/2005/D1.4.2/v3
Class Deliverable:	KWEB EU-IST-2004-507482
Version:	3
Date:	August, 2, 2005
State:	Final Version
Distribution:	Public

Copyright © 2005 The contributors

Knowledge Web Consortium

This document is part of a research project funded by the IST Program of the Commission of the European Communities as project number IST-2004-507482.

University of Innsbruck (UIBK) – Coordinator
Institute of Computer Science,
Technikerstrasse 13
A-6020 Innsbruck
Austria
Contact person: Dieter Fensel
E-mail address: dieter.fensel@uibk.ac.at

France Telecom (FT)
4 Rue du Clos Courtel
35512 Cesson Sévigné
France. PO Box 91226
Contact person : Alain Leger
E-mail address: alain.leger@rd.francetelecom.com

Free University of Bozen-Bolzano (FUB)
Piazza Domenicani 3
39100 Bolzano
Italy
Contact person: Enrico Franconi
E-mail address: franconi@inf.unibz.it

Centre for Research and Technology Hellas / Informatics and Telematics Institute (ITI-CERTH)
1st km Thermi – Panorama road
57001 Thermi-Thessaloniki
Greece. Po Box 361
Contact person: Michael G. Strintzis
E-mail address: strintzi@iti.gr

National University of Ireland Galway (NUIG)
National University of Ireland
Science and Technology Building
University Road
Galway
Ireland
Contact person: Christoph Bussler
E-mail address: chris.bussler@deri.ie

Universidad Politécnica de Madrid (UPM)
Campus de Montegancedo sn
28660 Boadilla del Monte
Spain
Contact person: Asunción Gómez Pérez
E-mail address: asun@fi.upm.es

École Polythechnique Fédérale de Lausanne (EPFL)
Computer Science Department
Swiss Federal Institute of Technology
IN (Ecublens), CH-1015 Lausanne.
Switzerland
Contact person: Boi Faltings
E-mail address: boi.faltings@epfl.ch

Freie Universität Berlin (FU Berlin)
Takustrasse, 9
14195 Berlin
Germany
Contact person: Robert Tolksdorf
E-mail address: tolk@inf.fu-berlin.de

Institut National de Recherche en Informatique et en Automatique (INRIA)
ZIRST - 655 avenue de l'Europe - Montbonnot Saint Martin
38334 Saint-Ismier
France
Contact person: Jérôme Euzenat
E-mail address: Jerome.Euzenat@inrialpes.fr

Learning Lab Lower Saxony (L3S)
Expo Plaza 1
30539 Hannover
Germany
Contact person: Wolfgang Nejdl
E-mail address: nejdl@learninglab.de

The Open University (OU)
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA
United Kingdom.
Contact person: Enrico Motta
E-mail address: e.motta@open.ac.uk

University of Karlsruhe (UKARL)
Institut für Angewandte Informatik und Formale Beschreibungsverfahren – AIFB
Universität Karlsruhe
D-76128 Karlsruhe
Germany

Contact person: Rudi Studer
E-mail address: studer@aifb.uni-karlsruhe.de

University of Liverpool (UniLiv)
Chadwick Building, Peach Street
L697ZF Liverpool
United Kingdom
Contact person: Michael Wooldridge
E-mail address: M.J.Wooldridge@csc.liv.ac.uk

University of Manchester (UoM)
Room 2.32. Kilburn Building, Department of Computer Science, University of Manchester, Oxford Road
Manchester, M13 9PL
United Kingdom
Contact person: Carole Goble
E-mail address: carole@cs.man.ac.uk

University of Sheffield (USFD)
Regent Court, 211 Portobello street
S14DP Sheffield
United Kingdom
Contact person: Hamish Cunningham
E-mail address: hamish@dcs.shef.ac.uk

University of Trento (UniTn)
Via Sommarive 14
38050 Trento
Italy
Contact person: Fausto Giunchiglia
E-mail address: fausto@dit.unitn.it

Vrije Universiteit Amsterdam (VUA)
De Boelelaan 1081a
1081HV. Amsterdam
The Netherlands
Contact person: Frank van Harmelen
E-mail address: Frank.van.Harmelen@cs.vu.nl

Vrije Universiteit Brussel (VUB)
Pleinlaan 2, Building G10
1050 Brussels
Belgium
Contact person: Robert Meersman
E-mail address: robert.meersman@vub.ac.be

Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document, even if they might not have directly contributed writing parts of this document:

University of Trento
France Telecom
University of Innsbruck
University of Sheffield
National University of Ireland Galway
University of Manchester
University of Liverpool
The Open University

Changes

Version	Date	Author	Changes
1	May, 28, 2005	L.Lancieri	Initial draft
2	June, 30 2005	L.Lancieri, D.Maynard, F.Gandon	First version (deliverable)
3	August, 2, 2005	L.Lancieri, D.Maynard, F.Gandon	Final Version

Executive Summary

The goal of this document is to make a synthesis regarding the practices of semantic web technologies. Our approach is oriented toward 3 main complementary directions.

The first aims at providing feedback on the opinions and the feeling of a group practitioners and researchers of semantic web technologies. The methodology is based on a "semi-closed" multiple choice online questionnaire. The results allow us to estimate the level of consensus of the community on concepts, methodologies and practices related to the Semantic Web.

The second direction aims to survey present and past initiatives related to best practices. Basically, this part is a synthesis of the effort of the W3C SWBP (Semantic Web Best Practices) that have a major involvement in this field.

In the third part, we investigate the practices in concrete developments. The field we studied is related to human language technologies. Starting from 4 case studies, we extract hints of useful practices that could help in the context of the Semantic Web.

This combined approach represents a first contribution of the Knowledge Web project to discussion regarding best practices. We shall see that, even if it is not always possible to find common approaches or a consensus, there are several proposals to clarify the practices. One of our major recommendations is to accentuate the effort in education initiatives in order to make Semantic Web technologies usable for a large majority. This deliverable aims at contributing to this effort.

Contents

1 Introduction

From a general point of view, the idea of collecting best practices starts from the need to have sufficient practical experience. This experience allows us to highlight consensus on positive and negative practices.

Regarding Semantic Web technologies, even if a large number of applications has begun to emerge, there are still many contradictory opinions. These opinions can be at the level of very basic concepts such as the actual usability of the semantic web, or they may be on specific technical points.

Since best practices means consensus, it is difficult to extract such a common point of view at this stage of the semantic web development.

In order to give some tracks of thinking for future investigations, we proposed an approach oriented in 3 main directions.

First of all, we created a multiple choice questionnaire that integrates frequent interrogations and possible answers. We adapted a methodology extracted from the field of collaborative work that enables to limit biases and optimise the statistic representativeness of answers. This methodology, described below, is certainly not perfect; however, the goal is not to obtain a definite opinion on best practices but to reduce the space of the problem. Even if this approach is a first step, the preliminary results show that consensus can emerge in some cases.

This feedback would certainly highlight the initiatives regarding guidelines and best practices. This is the next point we investigate. From these various initiatives; one of the more important is certainly the W3C SWBP (Semantic Web Best Practices).

From the users' feedback provided by the questionnaire and the recommendations from a standard body such as the W3C, it is now interesting to consider concrete technical practices. The field we studied is related to human language technologies. Starting from 4 case studies, we extract hints of useful practices that could help in the context of the Semantic Web.

We believe that these 3 complementary aspects -- feedback, recommendations and surveying concrete experience -- will contribute to providing useful and realistic advice to the industry.

2 From practices to best practices: definitions and introductive discussions

It appears that the boundaries of the notion of best practices can be very large and different from one expert to another. The list below seems to be a consensus. We try to use this W3C point of view as a reference.

2.1 Definition

2.1.1 From the W3C point of view

The W3C SWBPD working group (Semantic Web Best Practices and Deployment) defines the best practices as:

"A consensus-based guidance designed to facilitate Semantic Web deployment within RDF and OWL".

This includes:

Ontology representation practices:

- How to represent common ontologies (Unit, measure, etc);
- How to transform an existing representation into RDF/OWL representation;
- Interoperability with other external technologies (MPEG, UML, ..);
- Naming conventions for classes / properties / individuals;

Ontology engineering guidelines:

- How to design patterns for constructing ontologies;
- Ontology mapping (how to use multiples heterogeneous ontologies), mapping XML schema to OWL, integration, unification of ontologies;
- Practical deployment recommendations, guidance on how to "implement" the semantic web uses cases;

Software engineering guidelines for the Semantic Web:

- How to develop tools for the Semantic Web;
- How to develop tools that manage ontologies.

2.1.2 Use cases, Benchmarking and Best practices: What is a "good" practice?

Benchmarking seems a good approach to analyse most current practices in tools, architectures, methodologies, deployment etc (see the above definition). A well designed benchmarking evaluation can also lead to conclusions about whether practices are good or not. Indeed, benchmarking can contribute to evaluatinge the frequency of some practices and to estimate their efficiency. We postulate that with no contrary evidence, we can assume that the popularity is a clue of quality. A similar strategy can also be applied to the extraction of best practices from use cases.

The general idea is to identify, explain and disseminate: i) a good practice as: a practice that is considered good for the majority of experts or most frequent practices; ii) a bad practice as: a practice that is considered bad for the majority of experts.

3 What is the opinion of researchers and practitioners of Semantic Web technologies?

This section reports the methodology and the results of the online opinion poll.

3.1 Methodology

The questionnaire was developed in several stages. At each stage, the comments of several Knowledge Web members helped us to improve the content (correct / add questions and possible answers, etc.) and the ergonomic aspect of the questionnaire. After the first stage of improvement, the questionnaire was submitted to the industry area members and finally to all the project members. The idea is to enlarge this tested population in order to get the most realistic point of view.

The opinion poll is based on a semi-closed questionnaire that suggests responses but keeps the user free to make personal comments. The main advantage of such a questionnaire is to help extract major tendencies. We believe that even if there is a risk of biased results, such a methodology should be adopted (``the least bad solution``) when the question to be answered is not strictly defined or is subject to a large number of contradictory opinions. Providing an ``open`` questionnaire (i.e. where there are no proposed answers) could certainly be difficult to exploit, due to a lack of statistical representativness (too few answers, and therefore difficult to make a synthesis) and answers covering too large a spectrum (too many divergences).

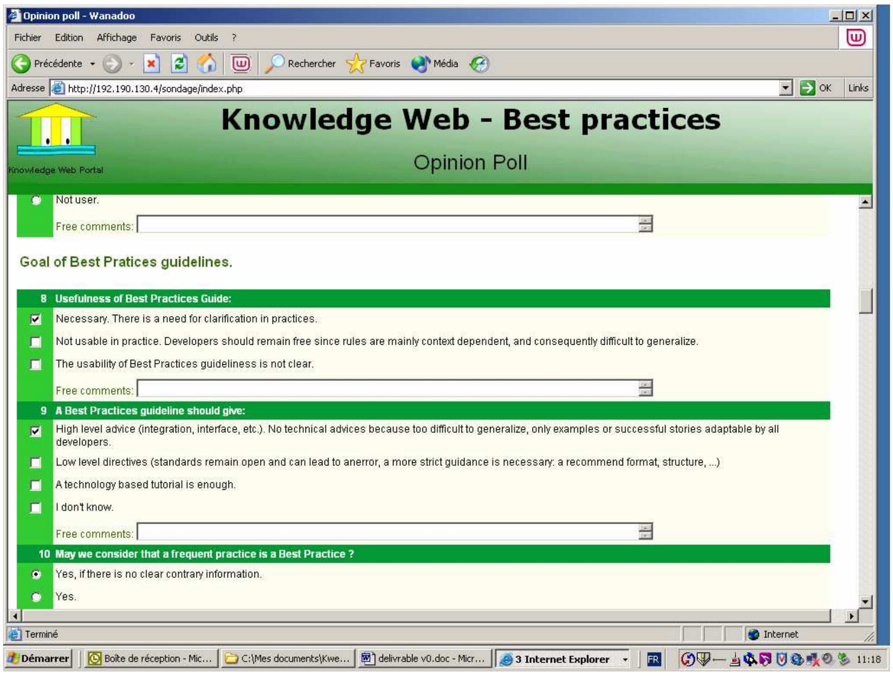
From the beginning of May 2005, the questionnaire has been publicly available online from the main Knowledge Web site and directly from the following URL:
http://192.190.130.4/sondage/index.php

We split the 38 questions into the following 7 categories reflecting a segmentation of the questions related to the semantic web:

- categorisation of the respondent (7 questions);
- goal of Best Practice guidelines (4 questions);
- ontologies and the real world (4 questions);
- building ontologies (9 questions);
- availability of ontologies (8 questions)l
- using ontologies (3 questions);
- technical concerns (3 questions).

These categories are detailed in the result section.

The following 2 figures show the answer interface. We can see for each question the proposed answers and the free area for personal comments. The second figure shows for each question the result area with a histogram presenting the repartition of the opinions and all the individuals' comments. Not shown here, at the end of the result area, are all the global comments. All this information can be consulted online.



Knowledge Web - Best practices

Opinion Poll

○ Not user.

Free comments:

Goal of Best Pratices guidelines.

8 Usefulness of Best Practices Guide:

☑ Necessary. There is a need for clarification in practices.

□ Not usable in practice. Developers should remain free since rules are mainly context dependent, and consequently difficult to generalize.

□ The usability of Best Practices guideliness is not clear.

Free comments:

9 A Best Practices guideline should give:

☑ High level advice (integration, interface, etc.). No technical advices because too difficult to generalize, only examples or successful stories adaptable by all developers.

□ Low level directives (standards remain open and can lead to anerror, a more strict guidance is necessary: a recommend format, structure, ...)

□ A technology based tutorial is enough.

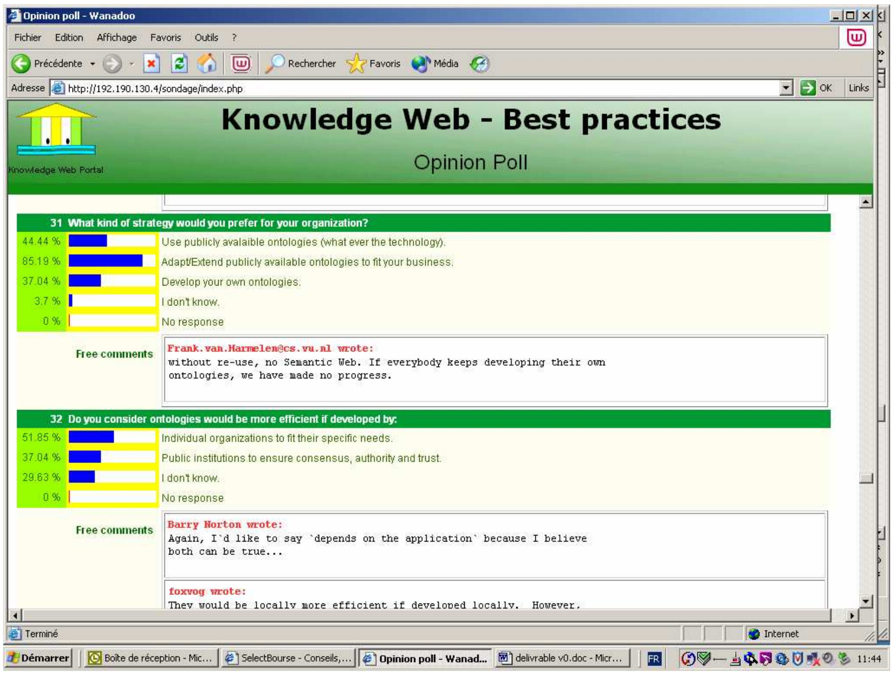
□ I don't know.

Free comments:

10 May we consider that a frequent practice is a Best Practice ?

● Yes, if there is no clear contrary information.

○ Yes.


Opinion poll - Wanadoo
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse http://192.190.130.4/sondage/index.php
OK
Links
Knowledge Web Portal
Terminé
Internet
Démarrer
Boîte de réception - Mic...
SelectBourse - Conseils,...
Opinion poll - Wanad...
delivrable v0.doc - Micr...
FR
11:44

Knowledge Web - Best practices

Opinion Poll

31 What kind of strategy would you prefer for your organization?

44.44 %	Use publicly avalaible ontologies (what ever the technology).
85.19 %	Adapt/Extend publicly available ontologies to fit your business.
37.04 %	Develop your own ontologies.
3.7 %	I don't know.
0 %	No response

Free comments

Frank.van.Harmelen@cs.vu.nl wrote:
without re-use, no Semantic Web. If everybody keeps developing their own ontologies, we have made no progress.

32 Do you consider ontologies would be more efficient if developed by:

51.85 %	Individual organizations to fit their specific needs.
37.04 %	Public institutions to ensure consensus, authority and trust.
29.63 %	I don't know.
0 %	No response

Free comments

Barry Norton wrote:
Again, I`d like to say `depends on the application` because I believe both can be true...

foxvog wrote:
They would be locally more efficient if developed locally. However.

Depending on the case, some questions need an exclusive answer (e.g. do you ever use OWL: yes or no) whereas others can accept several answers (e.g in what application do you use ontology?: information search, e-business, etc) . In this last case, the total percentage can of course be higher than 100%.

3.2 Results

In this section, we present the 7 categories of questions and a synthesis of the answers. This deliverable presents a static picture (30 contributors at the end of May 2005) of the community opinion. However, the detailed results are available online in real time, and therefore may have new contributors not present at the time of writing.

Presently the population of contributors is relatively small and the actual results should be considered with care because of the consequent low statistical representativeness. However, as mentioned previously, this stage of the work is intended to show trends and the questionnaire is planned to stay online during the entire project duration. One of the interesting future results could be the evolution of opinions throughout this period.

3.2.1 Who answered the opinion poll

This section is intended to provide information about the profile of the respondents, such as their professional occupation, category of institution and level of expertise in the technologies related to the semantic web.

The population is mainly composed of researchers (65%) including computer science (56%) and management study (7.5%). The other main part of the population (40%) is practitioners (professional developers, administrative, etc.). We can also remark that some respondents can have multiple profiles (e.g. computer science and linguist, etc).This is the reason why we get a total percentage higher than 100% (see section 3.1) The majority of the population also comes from academic institutions.

The self evaluation of the respondents regarding their knowledge and experience on the main tools and languages of the semantic web reveals that not many people consider themselves to have a large amount of knowledge. For XML 52% consider themselves advanced to expert, but for OWL this number is only 32% and for RDF it is only 30%. The levels of novice or non-user are 10 % (XML), 33 % (RDF) and 37 % (OWL).

Even if the tested population seems to have a fair knowledge of what the semantic web is, a high proportion (60 %) still feels that the concept is fuzzy or difficult to use, even if no one thinks that the concept is completely unusable. These results clearly show that an effort on education remains necessary.

These profiles can also be helpful to filter the final results. For example, we might want to compare the opinions of industrial practitioners with academic ones. It is also clear that it would be very interesting to compare the opinion of contributors having a good knowledge of RDF/OWL with the global population. Due to the limited amount of contributors, we present here a synthesis of global answers. We hope to enhance this aspect in future versions.

3.2.2 Goal of best practice guidelines

This topic aimed at obtaining feedback on the feeling of contributors about the usefulness of best practice guidelines, as well as what these might contain.

The majority (70%) think that there is a need for a clarification in practices, and developing best practice guides seems to be a reasonable approach. In this case the majority (63%) think that best practices should only consider high level advice (integration, interface, etc) and should avoid technical aspects which are too detailed. Some remarks consistent with the observation made in the previous section concerned the need for education (i.e. better practices come first from better knowledge). For 11% of the contributors, the usability of best practice guidelines is not clear and a technical tutorial is considered sufficient.

Other interesting remarks considered that best practice guidelines could be extended depending on the area of use, and in some cases could also integrate both high level and low level directives. There is also a small majority (60%) who wish to promote "labelling" through a certification authority, and who consider that basic and easily adaptable examples are better than nothing. For others (37%) it is not a good idea to implement this yet because of a lack of maturity.

An interesting divergence appears on the question related to the link between best practices and frequent practices. While 52% of the contributors think that a frequent practice should not be systematically be considered as a good practice, 45% think the contrary. This divergence induces the question of how to recognise a best practice. If we consider that a practice is based on previous uses and that expertise is based on the use of a technology, then frequent practices should be considered carefully, at least to start a recommendation repository. Alternatively we could consider that practitioners of a technology may also be influenced by bad habits coming from a "quick and dirty" adaptation of a theoretic principle. In this case frequent practices are not always good practices and "external" opinions coming from a recommendation group could be useful. Evidence for one expert is not necessarily evidence for another.

Other remarks pointed out that even if a frequent practice can provide a clue towards best practices, there is a need for more detailed technical, usage based advice or examples in order to be pedagogically useful.

3.2.3 Ontology and the real world

This section relates to the level of realism that ontologies should achieve. The question could be formulated in the following way: do we need practical concepts and tools which are easy to use if they only reflect poorly the real world, or should we instead promote precision in knowledge representation at the risk of introducing complexity?

Regarding the involvement of philosophers in semantic web, only 22% think that this is not a good idea (lack of pragmatism, difficult to manage, etc.) whereas 18% are clearly favourable. Actually the majority (60%) is mostly undecided and thinks that it should dependent on the context and application. The ratio is quite similar regarding the involvement of logicians. On the other hand, it seems that the help of linguists is a little more appreciated, since only 4% of the contributors think that a linguist would not be useful, whereas 33% are favourable and 70% think that it depends on the context and the application.

Uncertainty is linked to our perception of reality, and it is well known that our natural cognitive processes are mainly based on probabilistic reasoning. It could be interesting to ask whether uncertainty and probability need to be taken into account in the semantic web. The majority (67%) of the contributors answer yes to this question. The comments also clearly show that the semantic web is not mature enough to take into account these aspects.

3.2.4 Building ontology

Following from the previous question, we consider here practical aspects of ontology building.

Several respondents pointed out that RDF is very limited and cannot alone ensure the needs of the semantic web. Only 37% think that RDF alone could be enough, whereas 70% think that RDF and OWL are enough. 47% of the contributors prefer the use of a limited version of OWL (Lite, DL) instead of OWL Full. 37% think that embedding RDF in another technology (HTML, RSS, etc) should be recommended, whereas 26% recommend avoiding it (see details of the technical concerns in the questionnaire).

The majority (56%) of the contributors think that a domain oriented ontology (fit to the problem to be solved) should be recommended, whereas 30% think that a general ontology (a portable ontology usable in a maximum number of domains) is preferable, and 33% think that no rules should be recommended in this matter. Comments pointed out that the best way is probably to promote a domain oriented ontology linked to a general ontology.

The majority (78%) of the contributors think that the quantity of concepts used in a semantic web application should remain free since it depends on the application. About 10% think that there is a need for a maximum limit in order to reduce the complexity, possible inconsistency or to maintain good performance within the application.

For the majority (80%), the security aspects of an ontology mainly depend on the needs and context and it is difficult to be formalise these in strict rules.

The majority (67%) think that we need to recommend the use of ontology building from text (15% do not agree), such as tools for cleaning ontologies (59%) and consistency verification tools (74% in all cases, 19% only in complex cases).

Most contributors (41%) employ an ontology using only one natural language, whereas 26% use 2 or more. Regarding the representation language, 19% use one language whereas 19% use two and 19% use more than two. 40% of the contributors use synonyms for keywords, while 22 % do not.

3.2.5 Availability and reusability of ontologies

In order to improve the reusability of ontologies, we may wonder how to manage their availability. This includes preliminary considerations like persistency (i.e. building ontologies to be reusable, live for a long time, etc) but also the strategy of institutions (whether an ontology is freely available, etc.).

The majority of contributors (85%) think that an ontology is supposed to be persistent for a long time and can be used for several generations of applications. In such a case, a dedicated maintenance effort is necessary. Respondents also pointed out that this could depend on the context and that in some cases an ontology could have a limited time to live.

Strangely, only 18% of the contributors are sure that the semantic web will reach a high level of reusability, whereas 30% think that reusability will be low and 48% hope that the reusability will be high but that it is not clear that this will be the case. One respondent pointed out the need for popularisation of the ontology "model" (well modularised, easy to use, etc.).

Regarding the reuse of existing conceptualisations (database schemas, text, etc.), 48% of the contributors think that this should be promoted whereas 4% think the contrary and 44% think that it depends on the application. As suggested by some remarks, it is possible that the conceptualisation that the semantic web will ultimately be based on is not yet known. In such cases of conceptualization evolution, reuse of existing conceptualisations is certainly a need.

A majority is favourable to a mapping between new and existing ontologies (as a priority, 48%; if there is a need, 37%). The results show that reusability is a real concern within the semantic web community. Thus, 85% are considering adapting or extending an available ontology to their projects, whereas 37% prefer to develop their own ontology. The big discussion and opposition between specificity / optimality and openness / reusability appears again, considering that 52% think that an ontology would be more efficient if developed by an individual organization to fit their specific needs, whereas 37% think that this would be more efficient if done by a public institution in order to ensure authority, consensus, and trust. 30% do not have a clear idea on this subject.

Regarding availability, 37% think that ontologies should be available publicly, free of restrictions, whereas 48% think that it depends on the applications and that they could in certain cases be released under licence.

3.2.6 Using ontologies

This section is intended to give a feed back on the main uses of ontologies. The idea is to evaluate the level of applications where knowledge formalism is involved in machine to machine cooperation.

The results show that ontologies are used in a wide variety of applications; some (67%) are still mainly related to human-machine interaction (help with information search, browsing, etc.) whereas 63% are mainly inter-process related. The use of ontologies in e-business is 44%, but seems very promising as well as information disclosure and information integration. At the moment, security concerns do not seem to be a priority and few are taken into account in applications.

3.3 Discussion

The opinion poll is an interesting method of clarifying the perception of semantic web practices but, as mentioned in the introduction, this work needs to be improved.
The questionnaire focuses on the opinions of those contributors having a good experience in ontology building and manipulation. The methodology of our approach allows for example to filter the answers according to the level of experience in RDF/OWL (example of basic criteria). Such a focus will certainly improve the reliability of the conclusions drawn. This work needs to have more contributors (only 27 compared to the 170 Knowledge Web project contributors), and we hope to obtain this in the near future.

Thanks to the initial group of contributors, the questions and possible answers were clarified. This work allowed us to put online a first version of the questionnaire, although this also needs to be improved. New questions could certainly be identified, other clarified or eliminated.

In order to promote widely the results of such an opinion poll and encourage respondents, the ergonomics of the result presentation could certainly be enhanced. At the moment, only basic histograms are displayed. This could be sufficient to have a global feedback, but could be improved.

4 Survey of the activities in the W3C Semantic Web Best Practices and Deployment Working Group

4.1 Introduction

The aim of this W3C Semantic Web Best Practices and Deployment Working Group[1] (SWBPD) is to provide hands-on support for developers of Semantic Web applications. This working group helps application developers by providing them with "best practices" in various forms, ranging from engineering guidelines, ontology repositories to educational material and demo applications. The working group achieves its work through a mailing list, bi-weekly teleconferences on Mondays and by yearly face-to-face meetings. The activity of the working group is broken down into a number of task forces. Each of the following sections provides an overview of the work achieved so far by the different task forces.

4.2 OEP: Ontology Engineering and Patterns

The aim of the Ontology Engineering and Patterns task force[2] (OEP) is to provide guidance for developers of Semantic Web applications. In particular, OEP focuses on the

[1] http://www.w3.org/2001/sw/BestPractices/
[2] http://www.w3.org/2001/sw/BestPractices/OEP/

engineering of semantic web ontologies, through the publication of notes that document common and reusable ontology patterns, and general ontology engineering best practices. OEP tries, as much as possible, to:

- avoid judgments (good/bad) and concentrate on consequences of decisions and tradeoffs;
- avoid judgement calls and take specific issues, identifying representation/modelling choices;
- explain the consequences of choices, without claiming that they are"bad" or "good".

OEP produced two notes, and a number of drafts are being worked on:

- *Representing Classes As Property Values on the Semantic Web*[3] is a W3C Working Group Note since 5 April 2005. (Editor: Natasha Noy; Contributors: Michael Uschold, Chris Welty). The note addresses the issue of using classes as property values in OWL and RDF Schema. It is often convenient to put a class (e.g. Lion) as a property value (e.g. book subject) when building an ontology. The note presents various alternative mechanisms for representing the required information in OWL DL and OWL Lite: Approach 1: use classes directly as property values; Approach 2: create special instances of the class to be used as property values; Approach 3: create a parallel hierarchy of instances as property values; Approach 4: create a special restriction in lieu of using a specific value; Approach 5: use classes directly as annotation property values. For each approach, the note discusses various considerations that the users should keep in mind when choosing the best approach for their purposes.
- *Representing Specified Values in OWL: "value partitions" and "value sets"*[4] is a W3C Working Group Note since 17 May 2005 (Editors: Alan Rector) Modelling various descriptive "features", "qualities", "attributes" or "modifiers" is a frequent requirement when creating ontologies. For example "eye colour" may be constrained to take the values "blue", "green", "brown" or "black". In OWL, such descriptive features are modelled as properties whose range specifies the constraints on the values that the property can take on. This note describes two methods to represent such features and their specified values: 1) as partitions of classes; and 2) as enumerations of individuals. It does not discuss the use of datatypes to represent lists of values.
- *Defining N-ary Relations on the Semantic Web: Use With Individuals*[5] is a working draft (Editors: Natasha Noy, Alan Rector). In Semantic Web languages, such as RDF and OWL, a property is a binary relation; that is, it links two individuals or an individual and a value. This draft note presents ontology patterns for representing n-ary relations i.e. relations among more than two individuals or properties of a relation, such as severity or strength of a relation.

[3] http://www.w3.org/TR/2005/NOTE-swbp-classes-as-values-20050405/

[4] http://www.w3.org/TR/2005/NOTE-swbp-specified-values-20050517/

[5] http://www.w3.org/TR/2004/WD-swbp-n-aryRelations-20040721/

- *Simple part-whole relations in OWL Ontologies*[6] is an editor's draft (Editors: Alan Rector, Chris Welty). Representing part-whole relations is a very common issue for those developing ontologies for the Semantic Web. OWL does not provide any built-in primitives for part-whole relations (as it does for the subclass relation), but contains sufficient expressive power to capture most, but not all, of the common cases. The study of part-whole relations – mereology - is an entire field in itself: this note is intended only to deal with straightforward cases for defining classes involving part-whole relations. So far the note proposes 4 patterns.
- *Qualified cardinality restrictions (QCRs): constraining the number of property values of a particular type* is an editor's draft (Editors: Guus Schreiber). Cardinality restrictions are commonly used to constrain the number of values of a particular property, irrespective of the value type (e.g. hasCourse has a cardinality of 3 for a dinner). Sometimes we also need a way of saying that the number of values of a particular type (e.g. a starter) is restricted (e.g. to 1). We call these "qualified cardinality restrictions", where the term "qualified" means that we do not express restrictions on the overall number of values of a property, but only on the number of values of a certain type (i.e. class, datatype). So far the draft note proposes 3 approaches.

A few other topics are being considered for future drafts: semantic integration, fluents, units and measure, time and space, numeric range, etc.

4.3 PORT: Porting Thesaurii to RDF and OWL

The task force for Porting Thesaurii to RDF and OWL[7] is in support of the group's chartered aim of supporting the deployment in RDF/OWL of thesaurus (and similar) structured vocabularies. It has two short-term objectives: (1) a W3C Note on thesaurus and related techniques for the Semantic Web and (2) an RDF/OWL vocabulary for representing thesauri structures ('broader term' etc.) within RDF. In the longer term it is interested in:

- Document strategies for representing thesaurus-like content using RDF/OWL: produce guidelines for transforming an existing thesaurus (or classification system, or similar concept-based taxonomy) into an RDF/OWL representation. Guidelines should describe strategies for converting into an RDF representation of thesaurus-like structures, as well as strategies for re-describing in RDF/OWL the content originally conveyed in the thesaurus.
- Providing links to tools, applications, papers on this topic: the WG should seek to avoid duplicating existing work, and should provide links to existing efforts, encouraging feedback from implementers on the pros and cons of the approaches explored.
- Encourage dialogue between RDF and Semantic Web developers and members of the digital library community: many existing researchers in the digital library community (including Dublin Core and related) are using classification schemes

[6] http://www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole/index.html
[7] http://www.w3.org/2004/03/thes-tf/mission

and thesauri, and are not yet familiar with the facilities offered by RDFS and OWL. It is important to engage these communities rather than offer them solutions couched in the language of RDFS and OWL. In particular, concepts from the thesaurus world, such as 'facets', relate in non-obvious ways to similar, but more formalised, mechanisms offered by W3C's OWL technology.

This task force is currently focusing on two working drafts on SKOS. SKOS stands for Simple Knowledge Organisation System. The name SKOS was chosen to emphasise the goal of providing a simple yet powerful model for expressing knowledge organisation systems in a machine-understandable way, within the framework of the Semantic Web.

- *SKOS Core Vocabulary Specification*[8] a Working Draft (Editors: Miles, Brickley). SKOS Core is a model for expressing the structure and content of concept schemes (thesauri, classification schemes, subject heading lists, taxonomies, terminologies, glossaries and other types of controlled vocabulary). The SKOS Core Vocabulary is an application of the Resource Description Framework (RDF) that can be used to express a concept scheme as an RDF graph. Using RDF allows data to be linked to and/or merged with other RDF data by semantic web applications. A formal representation of the SKOS Core Vocabulary[9] is maintained in RDF/OWL.
- *SKOS Core Guide* a Working Draft (Editors: Miles, Brickley). This is a guide for using the SKOS Core Vocabulary, intended for readers who already have a basic understanding of RDF concepts.

4.4 VM: Vocabulary Management

Metadata element sets, taxonomies, subject headings, thesauri, and ontologies are all examples of vocabularies which are increasingly used in a "Semantic Web" environment. Managing vocabularies for use in Semantic Web applications means identifying, documenting, and publishing vocabulary terms in ways that facilitate their citation and re-use in a wide range of applications. The Vocabulary Management[10] task force examines practices in the maintenance communities for representative vocabularies ranging from small and informal to large and complex. It formulates principles of good practice and summarises discussion on issues for which good practice has yet to emerge. The task force identified several objectives:

- To establish the terminology for discussions of the declaration, identification, use, and management of vocabulary terms in a Semantic Web environment i.e. to list and define terms such as Term, Vocabulary, and Namespace.
- To articulate assumptions regarding the use of terms in a Semantic Web environment.
- To articulate guidelines of good practice for Namespace Owners to identify and declare Terms and Term Sets (Vocabularies) for use in a Semantic Web environment. Starting with fundamental guidelines such as "Identify Terms using

[8] http://www.w3.org/TR/swbp-skos-core-spec/

[9] http://www.w3.org/2004/02/skos/core

[10] http://www.w3.org/2001/sw/BestPractices/VM/

URIs", this section should formulate good-practice advice in areas where a workable consensus has developed on topics such as the backwards and forwards compatibility of URI-identified terms; the documentation of terms; "namespace" policies; "ownership" of namespaces; and approaches to versioning terms and identifying term versions.

- To point to and briefly summarize the evolving diversity of practices and ongoing approaches to declaring and managing vocabularies. Examples are the question of what sort of human-readable or machine-processable documents, if any, term URIs should "resolve to"; how an organisation or even an individual can go about declaring and publishing a term or a vocabulary; and how "good" URIs should be formed.

There is a draft on the W3C Wiki of the note[11]. The headers of the working draft include:

- Identify Terms with URI References,
- Articulate and publish maintenance policies for the Terms and their URI references,
- Identify the historical version of a Vocabulary or its Terms,
- Provide natural-language documentation about the Terms,
- Declare the Terms using a formal, machine-processable schema language,
- What should the identifier of a Vocabulary or Term resolve to,
- What does it mean to "use" Terms from one Vocabulary in another,
- What does it mean to "own" a Vocabulary,
- When a term is needed, when should one adapt an existing term, declare a new one, or get an established vocabulary maintainer to host

4.5 XSCH: XML Schema Datatypes

The XML Schema Datatypes[12] task force considers two issues:

- what URI should be used within RDF and OWL for user defined XML Schema Datatypes;
- what is the relationship between the value spaces of the various XML Schema built-in simple types when used within RDF and OWL.

The working draft *XML Schema Datatypes in RDF and OWL*[13] (Editors: Jeremy J. Carroll, Jeff Z. Pan) explains that RDF and OWL Recommendations only use the simple types from XML Schema and discusses three questions left unanswered by these Recommendations:

- What URIref should be used to refer to a user defined datatype?
- Which values of which XML Schema simple types are the same?
- How to use the problematic xsd:duration in RDF and OWL?

The note also discusses the use of numeric types.

[11] http://esw.w3.org/topic/VocabManagementNote

[12] http://lists.w3.org/Archives/Public/public-swbp-wg/2004Apr/0125.html

[13] http://www.w3.org/TR/swbp-xsch-datatypes/

4.6 HTML: Embedding RDF in HTML

There is a long standing requirement to embed metadata in an HTML document. One would think that this requirement could be satisfied by combining XHTML with RDF/XML using XML Namespaces and XML Schema, but this is not so[14]. Instead, there are many nuanced technical issues and a series of problem statements, and the goals of this task force[15] are to identify the requirements and constraints for embedding RDF in (X)HTML and to document a solution for satisfying those requirements.

There are two documents being discussed:

- *RDF/A Syntax A collection of attributes for layering RDF on XML languages*[16], (Editors: Birbeck, Pemberton (eds.) a note since 11 October 2004 that outlines a syntax for layering RDF information on any XML document, via attributes.
- *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)*[17], (Editors: Hazaël-Massieux, Connolly) a W3C team submission since16 May 2005. GRDDL is a mechanism for Gleaning Resource Descriptions from Dialects of Languages; that is, for getting RDF data out of XML and XHTML documents using explicitly associated transformation algorithms, typically represented in XSLT.

4.7 ADTF: Applications and Demos

Until now this task force[18] maintained a weblog of Applications and Demos[19]. However the process of collecting them is slow and tedious and a new proposal below was designed to speed up the process, by persuading people to document the applications and demonstrators they make using RDF. (Note that this is a proposal only at the moment June 2005):

At the Boston face to face meeting of the working group the following proposal was minuted:

- select and make public the criteria for inclusion;
- write up some information about how to create a DOAP file for this purpose;
- continue to use the weblog to create links but use a link to DOAP file rather than GRDDLing out DOAP;
- encourage people to create DOAP files for their apps and demos;

4.8 RDFTM: RDF/Topic Maps Interoperability

The RDF/Topic Maps Interoperability task force[20] of the Semantic Web Best Practices and Deployment WG is in support of the group's chartered aim of providing guidelines

[14] http://www.w3.org/2003/03/rdf-in-xml.html
[15] http://lists.w3.org/Archives/Public/public-rdf-in-xhtml-tf/
[16] http://www.w3.org/MarkUp/2004/rdf-a.html
[17] http://www.w3.org/TeamSubmission/grddl/
[18] http://esw.w3.org/topic/SemanticWebBestPracticesTaskForceOnApplicationsAndDemos
[19] http://esw.w3.org/mt/esw/archives/cat_applications_and_demos.html
[20] http://www.w3.org/2001/sw/BestPractices/RDFTM/

for users who want to combine usage of the W3C's RDF/OWL family of specifications and the ISO's family of Topic Maps standards.

The short term objectives of the task force are to:

- Document strategies for representing topic maps using RDF/OWL and vice versa.
- Describe the pros and cons of existing approaches.
- Produce guidelines for transforming a topic map into an RDF/OWL representation and vice versa.
- Provide links to tools, applications, and papers on this topic.

Longer term objectives of the task force include:

- Proposing the guidelines described above for standardisation in the W3C and ISO.
- Producing guidelines for using OWL to constrain topic maps.
- Producing guidelines for cross-querying RDF/OWL data and topic maps.

The task force produced a working draft called *A Survey of RDF/Topic Maps Interoperability Proposals*[21] (Editors: Steve Pepper, Fabio Vitali, Lars Marius Garshol, Nicola Gessa, Valentina Presutti). This draft contains a survey of five proposals for integrating RDF and Topic Maps data and is intended to be a starting point for establishing standard guidelines for RDF/Topic Maps interoperability.

The task force also works on *Test Cases for RDF/TM Interoperability*[22].

4.9 SE: Software Engineering Task Force

The Software Engineering Task Force[23] investigates potential synergies between the Semantic Web and domains more traditionally associated with Software Engineering. This is to enable the promotion and cross-pollination of both new and established ideas between the two communities, potentially relating to:

- Use cases;
- The application of models, patterns and frameworks;
- Methods and tools;
- Underpinning technologies;
- Best practice.

Objectives include:

- To collect, collate and validate a list of potential ideas and uses for the Semantic Web in Software Engineering and to make this list publicly available,
- To further evaluate ideas already presented to the Semantic Web Best Practices Working Group. These include:
  - The potential for Ontology Driven Software Engineering, Ontology Driven Architectures (ODA) and the crossover between Ontology Engineering and Software Engineering;
  - The use of composite identification schemes on the Semantic Web and their potential use for 'ontology joining' and the reduction of ambiguity across the Software Lifecycle;

[21] http://www.w3.org/TR/2005/WD-rdftm-survey-20050329/

[22] http://tesi.fabio.web.cs.unibo.it/cgi-bin/twiki/bin/view/RDFTM/TestCases

[23] http://www.w3.org/2001/sw/BestPractices/SE/

- The construction of dynamic self-organising applications using Semantic Web technologies;
- The use of Semantic Web Technologies to produce highly adapted/adaptive (user) interfaces and support tools.

The first draft of this task force is called *Ontology Driven Architectures and Potential Uses of the Semantic Web in Software Engineering*[24] (Editors: Phil Tetlow, Jeff Pan, Daniel Oberle, Evan Wallace, Michael Uschold, Boeing, Elisa Kendall). It is considered by many that applying knowledge representation languages common to the Semantic Web, such as RDF and OWL, to Systems and Software Engineering can achieve significant benefits. This note hence attempts to outline such benefits and the approaches needed to achieve them from a Systems and Software Engineering (SSE) perspective. It is aimed at professional practitioners, tool vendors and academics with an interest in applying Semantic Web technologies in Systems and Software Engineering contexts.

Other interesting topics include a discussion on a note "from object-oriented design to semantic web modeling"[25]. A number of different problems were discussed[26] such as: OO subclass vs. OWL subsumption, how to characterise the notion of "the class used when an object was created" (like in OKBC that had the notion of "direct-type"), semantics of "slot attachment" in OO and domain and range restrictions, openness of OWL which is weird to someone from an object modeling background, etc. In addition it was suggested to have a special attention for UML users and JAVA users, to address their needs and if possible use terminology familiar to them.

4.10 WordNet Task Force

Wordnets are valuable resources both as lexical repositories and as sources of ontological distinctions. The WordNet Task Force[27] is in support of the group's chartered aim of supporting the deployment in RDF/OWL of WordNet and similarly structured lexica ("wordnets").

The main short-term objective is to document strategies, examples and resources for representing wordnet-like content using RDF/OWL: The task force should produce guidelines for transforming existing wordnets into an RDF/OWL representation. Guidelines should describe strategies for converting wordnet-like structures into an RDF representation, as well as strategies for re-describing in RDF/OWL the content originally conveyed in the wordnets. It should also recommend an RDF/OWL vocabulary for representing wordnet structures ('synset' etc.) within RDF

Many existing researchers in the lexical semantics community (including Princeton WordNet developers and related initiatives, see list at bottom of page) are using wordnets, and some are not yet familiar with the facilities offered by RDFS and OWL. It is important to engage these communities rather than offer them solutions couched in the language of RDFS and OWL. In particular, concepts from the wordnet world, such as

[24] http://www.w3.org/2001/sw/BestPractices/SE/ODA/

[25] http://lists.w3.org/Archives/Public/public-swbp-wg/2004Oct/0096.html

[26] http://lists.w3.org/Archives/Public/public-swbp-wg/2004Oct/0113.html

[27] http://www.w3.org/2001/sw/BestPractices/WNET/tf

'synsets' and 'hyperonymy' relate in non-obvious ways to similar, but more formalised, mechanisms offered by W3C's OWL technology.

Among the documents produced by this task force are:

- *Porting Wordnets to the Semantic Web*[28] an editor's draft 8 July 2004. This draft presents a framework and workplan for porting wordnets to Semantic Web languages, like RDFS and OWL. Some phases are distinguished, and preliminary resources are referenced.
- WordNet datamodel[29]
- Wordnet in RDFS and OWL[30]. This sketches a draft to describe an RDF Schema and OWL ontology for representing WordNet.

4.11 Semantic Web Tutorials

This really is a web page[31] that provides a central collection of Semantic Web tutorial resources for interested readers and is maintained by the Working Group.

FAQ - Frequently Asked Questions

Recently the working group started a new action to address the problem of navigation in the best practices. For instance a design pattern has one name but the problems it solves could be described in very different terms and thus its one name is not enough to index it. The working group is starting a FAQ system to provide a parallel indexing of the topic addressed in the different notes.

[28] http://www.w3.org/2001/sw/BestPractices/WNET/Porting

[29] http://www.w3.org/2001/sw/BestPractices/WNET/wordnet_datamodel.owl

[30] http://www.w3.org/2001/sw/BestPractices/WNET/wordnet-sw-20040713.html

[31] http://www.w3.org/2001/sw/BestPractices/Tutorials

5 Example of best practices and success stories in human language technologies

5.1 Introduction

In this section we describe some Human Language Technology applications for the Semantic Web which have been specifically designed for use in industrial settings and which have been implemented and tested. We discuss the motivation and need for such products in each case, and give details of the application and its performance or evaluation in a real-world setting. We describe 4 different systems, all of which are based on the underlying Information Extraction technology provided by the University of Sheffield's GATE [Cun02b], but which are used in very different ways. The first two systems, KIM and SWAN, are quite generic and designed for more general purpose use, SWAN's main selling point being its scalability. The third system, h-TechSight, is designed for more precise use in very specific domains (currently chemical engineering), although it can be adapted to different domains assuming the support of appropriate ontologies, and aims particularly to target and capture information which changes over time. Finally, Rich News is designed to work on general news texts, but in a rather specific way, in that it addresses the issue of enabling access to broadcast news.

5.2 KIM

5.2.1 Motivation

KIM is designed as a multi-purpose knowledge management (KM) platform, enabled to serve a wide variety of information needs and KM tasks in different domains and configurations. In essence, it allows for management of texts and ontologies in connection with each other. Its advantages compared with other contemporary information systems can be summarized as follows:

- it allows the combination of FTS (full-text search, i.e. the simplest form of IR) with structured (DB-like) queries. An example of this might be the case of asking for all documents which contain references to objects matching a structured query.
- The structured queries are performed on top of a semantic store. Based on automated interpretation (reasoning, inference) on top of the semantics of the data (the ontology), the semantic store is capable of answering questions based on data which has not be explicitly asserted. For example, it is capable of returning "Mary" as a result of the query <*John*, *relativeOf*, *?x*> based on an ontology of family relationships and the assertion <*Mary, motherOf, John*>.

- KIM is capable of analysing the documents automatically in order to populate the ontology and link the documents to the structured knowledge.

A business intelligence scenario for the use of KIM might consist of a set of texts (news, reports, etc) indexed with KIM, which would be able to match the query:
"*documents speaking of a telecom company in Europe and John Smith*" with a document containing "*The board of Vodafone appointed John G. Smith as a CTO*."
KIM's advantage is that a "regular" information retrieval system cannot match:

- Vodafone with "telecom in Europe" because it doesn't know that:
    - Vodafone is a mobile operator, which is a sort of telecom company;
    - Vodafone is in UK, which is part of Europe;
- "John G. Smith" with "John Smith".

5.2.2 Product

The KIM (Knowledge and Information Management) Platform [Pop04a] is an efficient, robust, and scalable architecture for automatic semantic annotation, implemented in a component-based platform for semantic-based indexing and retrieval from large document collections. KIM offers an end-to-end, extendable system which addresses the complete cycle of metadata creation, storage, and semantic-based search and includes a set of front-ends for online use, that offer semantically enhanced browsing.

The KIM platform consists of formal knowledge resources (KIM Ontology and instance/knowledge base), the KIM Server (with API for remote access, embedding, and integration), and front-ends (browser plug-in, KIM Web UI, and Knowledge Base Explorer). The architecture of the KIM Server allows for easy modification, extension, and embedding in third-party systems. It also provides an abstraction layer over the specific underlying component implementations, and thus ensures flexibility in cases of a custom implementation (or configuration) of KIM with another semantic repository, metadata storage or IR engine. The KIM Server has the following major components: Semantic Repository, Semantic Annotation, Document Persistence, Indexing and Query.

KIM contains an instance base which has been pre-populated with 200,000 entities of general importance that occur frequently in documents. The core entities are different kinds of locations: continents, countries, cities, etc. Each location has geographic coordinates and several aliases (usually including English, French, Spanish, and sometimes the local transcription of the location name) as well as co-positioning relations (e.g. subRegionOf). As previously shown by [Mik99b], IE systems need such data, because locations are difficult to recognize otherwise.

The information extraction in KIM is based on the GATE Framework [Cun02b]. The essence of the KIM IE is the recognition of named entities with respect to the KIM ontology (KIMO). The entity instances all bear unique identifiers that allow annotations to be linked both to the entity type and to the exact individual in the instance base. For new (previously unknown) entities, new identifiers are allocated and assigned; then minimal descriptions are added to the semantic repository. The annotations are kept separately from the content, and an API for their management is provided.

More information about KIM is described in the Knowledge Web Deliverable D1.2.2 SWF Requirements Analysis, where the interoperability between different semantic annotation systems and their components is investigated.

5.2.3 Discussion

For the end-user, KIM's Information Extraction functionality is straightforward and simple. The user requests information from a browser plug-in, which highlights the entities in the current content and generates a hyperlink used for further exploration of the available knowledge for the entity. Various access methods are also available, e.g. entity pattern search, entity lookup, keyword and document attribute search. There is also an opportunity to create a composite query consisting of atomic searches of the above types. This means that the product is very suitable for use by non-experts, because they do not require technical knowledge about how the system works in order to get results easily and in a very visual way.

The KIM Plugin is currently being evaluated in terms of usability as part of WP 1.2 Evaluation for Technology Selection. This work will be reported in the forthcoming Deliverable D1.2.1. Essentially, KIM scores very highly on factors such as ease of setup and use, documentation quality, and aesthetics, although there are some accessibility issues which still need to be resolved.

The performance of the Information Extraction component, which is the driving force behind KIM, has been measured against a human-annotated corpus containing 100 news articles from UK media sources. This corpus has been annotated with a flat structure of Named Entities (Person, Organisation, Location, Date, Percent, and Money). Although KIM recognises more specific information about entities than this, i.e. it attaches instances to subsuming nodes in the ontology, it can still be evaluated according to these more general concepts. Overall the system currently achieves an average of 86% Precision, 84% Recall and 85% F-Measure. Work is currently underway in Knowledge Web Work Packages WP2.1 and WP2.3 to develop more advanced evaluation metrics and software capable of evaluating named entities according to a full ontology rather than as a flat structure. We shall be evaluating KIM and other systems using these metrics and evaluation tools over the coming months.

5.3 SWAN

5.3.1 Motivation

The Semantic Web and Semantic Web Services represent the next stage of evolution for the web, distributed computing and collaborative science. Key to the success of this is the production and maintenance of formal data, in the form of ontologies and related instance sets or knowledge bases. Whereas the simplicity of HTML and the ubiquity of natural language led to the organic growth of the hypertext web, semantic data is much harder to create and maintain. Human Language Technology provides the missing link between natural language and formal data, thereby glueing together web services and their user constituency, and facilitating enterprise integration. There is currently much work in the area of semi- and fully automatic semantic annotation, but until now there has always been a tradeoff between performance and scalability. While performance is clearly important, the semantic web will never be a reality unless applications are fully scalable and can cope with enormous volumes of data. Systems that are designed for massive annotation are generally automatic, non-specific and do not have a high level of performance. Smaller, more targeted systems may perform well but are not scalable to large amounts of data.

5.3.2 Product

SWAN (Semantic Web ANnotator) is a system designed to perform large-scale ontology-based information extraction for the semantic web, annotating vast amounts of documents from the web with semantic information (inferred metadata). The annotation process can be viewed as a chain of logical components, starting with the crawling of documents from the web and ending with the user of the platform receiving a semantic response to a query. The system is based largely on KIM [Pop04a], which provides indexing, disambiguation and storage components, as well as some of the interface components.

SWAN contains two focused crawler versions: an HTML crawler which directly accesses web pages according to a defined scope, and an RSS crawler which uses the syndication mechanism of RSS 1.0 newsfeeds. The RSS crawler has the advantage of being already domain-specific and therefore more likely to return relevant documents, and some "free" (explicit) metadata such as author name and publication date. The web pages found are then passed to the IE component, which consists of a set of processing resources implemented using GATE [Cun02b]. This pipeline of resources performs preprocessing tasks such as tokenisation and sentence splitting, followed by high-level pattern matching and coreference resolution, and results in a set of semantic annotations linking the text with concepts from an ontology. The disambiguation component then performs 2 tasks: first, it co-refers different mentions of the same instance at the document level, and second, it continuously checks if new instances found are identical to previously found entities in other documents (and thus already contained in the Knowledge Repository). Finally, the results are stored in various databases. Entities, relations and their properties are stored in an RDF Knowledge Repository, using Sesame[32]. An index relating the entities to their source documents is stored in a Document Store, implemented on top of Lucene[33]. The annotations themselves are stored in an Annotation Store implemented as a relational database.

SWAN allows access to its data for humans via a web-based UI, using an ordinary web browser, which allows the user to enter queries, e.g. "Who are the CEOs of companies in Ireland?", and to access the results via a web page. They can also pose queries directly in a formal query language such as RQL or SeRQL, and access the results as RDF statements about the entities matching the query. The system is designed to work on specific domains, because the accuracy is vastly improved in this way. However, it is also deliberately designed to be scalable, and new domains are being continuously added.

5.3.3 Discussion

SWAN has been evaluated in a number of ways. The problem of scalability with respect to crawling and annotation is dealt with by organising the components in a cluster architecture of 4 annotator machines responsible for the extraction process. A document queueing system divides the load between the 4 machines. The crawler places each downloaded document on top of the queue, and each annotator in turn takes a document from the queue and processes it. An upper limit is set for the queue size to prevent overload -- if this limit is reached then the crawler halts temporarily. The number of machines could of course be increased, should the need arise. A distributed architecture has not been implemented for storage, but the current architecture appears to scale well in tests so far.

5.4 h-TechSight

5.4.1 Motivation

The growing pervasiveness of Knowledge Management (KM) in industry marks an important new watershed. KM has become embedded in the strategy, policy and implementation processes of institutions and organisations worldwide. The global KM market has doubled in size since 1991 and is projected to exceed US$8.8 billion in 2005. KM applications are expected to save Fortune 500 companies around $31 billion, and the

[32] http://www.openrdf.org
[33] http://lucene.apache.org

broader application cost has similar projected forecasts. Although the tools and resources developed in h-TechSight are targeted towards SMEs, there are important implications for the growth and dispersion of such new technologies to industry as a whole. h-TechSight aims to pave the way for such development by providing a variety of knowledge management tools in its portal.

The hTechSight Knowledge Management Portal (KMP) [May04b,Maynard05b] has two modes of use: a generic application which performs ontology-enhanced information retrieval facilities, and a targeted application which provides mechanisms for knowledge acquisition in specific domains. Currently it covers the employment and news domains within the field of Chemical Engineering.

Employment is a generic domain into which a great deal of effort in terms of knowledge management has been placed, because every company, organization and business unit must encounter it. Human Resources departments often have an eye open for knowledge management in order to monitor their environment in the best way, and many recruitment consultant companies have watchdogs to monitor and alert them to changes. There exist a variety of job search engines (portals) which use knowledge management extensively to link employees and employers, e.g. JobSearch[34] and Job Portals[35]. The employment application in the KMP aims to alert users to technological changes, since job advertisements are a very good indicator of moving trends in the field. By monitoring these advertisements over a period of months or even years, we can examine, for example, changes in the requirements for particular skills and kinds of expertise required, how salaries fluctuate, what kinds of qualifications are being demanded, and what kinds of benefits packages employees can expect.

The news domain is another clear area where it is important for companies to keep a close eye on technological developments in their field. Primary market players for this are the pharmaceutical industry and the oil and gas industry. Pharmaceutical companies need to extract knowledge from diverse sources in order to predict pharmacological and toxicological effects, for example integrating knowledge from newly acquired organisations and keeping a close watch on news of and reports from their competitors. The oil and gas industry is currently faced with increasing pressures to create higher quality and more environmentally friendly products, and therefore such companies need up-to-the-minute access to news, reports, and experiences of colleagues around the world in order to leverage such information and respond to critical information requests from government agencies. The application for the news domain is aimed at helping companies to access and monitor such information quickly and accurately, bringing new products, processes and technologies to their attention, as well as tracking the progress of rival companies in the field.

[34] http://www.job-search.com/

[35] http://www.aspanet.org/solutionstemp/jobport.html

5.4.2 Product

The h-TechSight KMP aims to extract new domain data from free text on the web. It uses GATE to power the concrete data-driven analysis of concepts and instances in the knowledge management platform, with respect to an ontology and domain. The GATE Information Extraction (IE) application enables statistical information to be gathered about the data collected, and inferences drawn, which in turn leads to the monitoring of trends of new and existing concepts and instances.

The application uses two main inputs: a web mining application which feeds relevant URLs to GATE, based on the user's query, and a domain ontology. The texts are automatically annotated with semantic information based on the concepts in the ontology. When an instance of a concept is found, it is annotated with semantic metadata. Instances in the text can not only be visualised (through colour-coding) but are also output in two forms: into a database for further processing, and in ontological form. On the one hand, this annotation of semantic metadata enriches the text; on the other hand, the ontology may be enriched through its population with instances from the text.

h-TechSight performs metadata generation and ontology population (by adding new instances to the ontology), but also by enabling the process of ontology *evolution*. By this we mean that the IE application serves not only to populate the ontology with instances, but also to modify and improve the ontology itself on the conceptual level. Statistical analysis of the data generated can be used to determine how and where this should take place. For example, a set of instances will be linked to a concept in the ontology, but this concept may be too general. A clustering algorithm can be used to group such instances into more fine-grained sets, and thereby lead to the addition of new subconcepts in the hierarchy. hTechSight is also unique in performing monitoring of the data over time, which can also lead to suggested changes in the ontology.

5.4.3 Discussion

The IE application has been evaluated in terms of Precision and Recall to see how well the system finds relevant instances of the concepts. The system was tested on a set of 38 documents containing job advertisements in the Chemical Engineering domain, mined from the website http://www.jobserve.com. The web portal was mined dynamically using a web content agent written in WebQL, a commercial web crawling software[36]. These documents were manually annotated with the concepts used in the application, ad the evaluation tools provided in GATE were used to compare the system results with the gold standard. Overall, the system achieved 97% Precision and 91.5% Recall, with an F-

[36] http://www.webql.com

Measure of 94.2%. This high level of performance is more than adequate for most users' needs, as in any case any erroneous results can be manually corrected.

The KMP has been tested by users in industry, such as Bayer Technology Services, JetOil and IChemE. Users found that it was very helpful in increasing the efficiency of acquiring knowledge and supporting project work in industry, by helping to automatically scan, filter, structure and store the wealth of information available on the web related to their needs. For Bayer, the potential areas of application spanned from research and development, engineering and production, to marketing and management.

Users at IChemE, a leading international body which provides services for chemical engineers world-wide, claimed that the employment application was a very sound idea, and that it "would be a very valuable means of graduates gaining a fresh insight into their jobs and related training which may be narrower than ideally it should be due to company constraints (i.e. time and money for development)".

5.5 Rich News

5.5.1 Motivation

Rich News seeks to address the problem of how to improve access to the large amounts of broadcast audio and visual material produced by media organizations. Material can only be effectively accessed if metadata describing it is available in some sort of cataloguing system. The British Broadcasting Corporation (BBC), who produced the material on which Rich News was developed, produce material for four television channels, nine network radio stations, and numerous local radio stations. Manual annotation of this material by an archivist is an expensive and labour-intensive task. For example, it takes a BBC archivist almost seven hours to catalog Newsnight, a fifty minute daily news broadcast, in detail. Because of the high cost of cataloging, 90% of the BBC's output is annotated only at a very basic level, making it difficult to re-use it after its initial broadcast. Furthermore, because of the time it takes for cataloguing to be completed, there is a delay before the material is available, which can be a problem in areas such as news and current affairs, when the material is most likely to be useful immediately after it is broadcast.

A system able to automate, or partly automate the annotation process is therefore very useful. While producing a system that annotates as accurately and with as much detail as a human annotator does is beyond the scope of present technology, it is clear that a system that provided less detailed and less reliable annotations would still be useful. With such a system, inaccuracies or omissions might prevent access to some material, or suggest that material was relevant when it was not. However, at present no annotations are produced manually for much broadcast output, preventing effective access to it, so a level of performance for such a system of well below 100% would be acceptable. In addition, the automatic linking of web and multimedia content enables a new model of mixed-mode media consumption [Dim04b].

Blinkx and Google have both recently launched television search engines, but those systems rely on a simple text-matching search, and do not use the inherent structure of broadcasts to aid in the retrieval process. Previous work has adopted similar information extraction technologies to those used here (see for example [Przybocki99]), but Rich News is novel in both the use of web-based content augmentation and in the use of semantic annotation.

Rich News therefore allows high quality textual and semantic metadata to be produced fully automatically for news broadcasts. The resulting annotations can be viewed together with the original media file in a multimedia annotator, thus allowing the annotations to be searched, manually corrected or enabling supplementary annotations to be added by an archivist. Rich News can then produce index documents for individual news stories, containing links to the recordings of the broadcasts in which they occur, as well as textual and semantic metadata. These can be searched using the Web User Interface of the KIM system.

5.5.2 Product

The Rich News system automatically annotates radio and television news broadcasts with textual content, using resources retrieved from the web. It identifies individual stories in news broadcasts and annotates them with related material, which is then semantically analysed and used to produce summary information for each news story.

The overall annotation system can be divided into seven sequential modules. The first four modules are a speech recognition module, a module that divides the broadcast into segments corresponding to individual news stories, a module that finds keywords for each story, and a module that finds web pages reporting the same story as that reported in the broadcast. At this stage the 6th module, manual annotation, may be undertaken. The penultimate module makes a story index document for each story in the broadcast, and the final module, KIM, performs information extraction and semantic annotation on the text of the web document, thus allowing the named entities in the broadcast story to be identified.

The annotation process starts by performing automatic speech recognition to achieve a rough transcript for each program, and then analysing this transcript to determine the boundaries between the various news stories that it describes. This task is made difficult due to errors in the output of current large vocabulary speech recognition systems. Rich News then tries to find keywords or phrases that describe the content of each story. Using these key phrases, and the date of the program, it is possible to search on the BBC website to find web pages that are likely to be related to the story. By downloading the candidate web pages and comparing their text to the transcript of the broadcast, it is usually possible to find a web page reporting the same news story that was in the broadcast. The section of the web page containing the news story can give us a classification for the story, which in some cases is quite detailed, such as the particular English county it relates to. Summaries and titles for the stories can also be extracted from the explicit metadata in the web pages.

Because the text in the web pages is error free, and contains useful cues such as capitalisation and punctuation that is missing from the transcripts, it is furthermore much easier to use this data as a basis for further analysis. The KIM information extraction system is therefore used to find entities in the web pages related to each story, and these are annotated with semantic classes, allowing the stories to be indexed and queried in much more flexible ways than if text search alone were used.

5.5.3 Discussion

The performance of the RichNews annotator is largely dependent on how successfully the annotator produces index documents, which in turn is dependent on how successful it is in finding webpages for the stories in the broadcasts. Therefore evaluation of the system is based on a measure for determining the proportion of news stories in broadcasts for which Rich News was able to produce appropriate index documents automatically.

Evaluation of the system's performance has been conducted by first playing nine broadcasts, and noting the stories that occurred in each. The programs used in the evaluation were from BBC Radio 4's The World at One (a 30 minute daily national news program), taken from the last six months of 2002. Once each story appearing in each broadcast had been noted, Rich News Annotator was run on each of the broadcasts, and story index documents were produced. For each index document, it was determined whether it reported a story covered in the corresponding radio broadcast, whether it reported a closely related story, but could not be said to be reporting a story in the broadcast, or whether it reported an unrelated story.

Results were calculated under two conditions. In the first condition, strict, annotation was only considered successful if the correct story was matched, but in the second, lenient, it was considered correct if a closely related story was matched. The nine broadcasts considered contained a total of 66 news stories. The results of the evaluation show that the system achieved very high precision of 92.6% (strict) and 100% (lenient), although the recall was somewhat lower (37.9% strict and 40.1% lenient).

The current state of the performance clearly demonstrates that Rich News Annotator, running in its fully automatic mode, can give access to a large volume of material that would be inaccessible if no annotation were provided, which is the case with much of the BBC's output at present. The system is currently in use, though work is still ongoing on the system, and efforts are being made to exploit the redundancy available in multiple news websites in order to improve recall. The stories that were missed by the annotator were often those that consisted of only one or two sentences, rather than those that were reported in more depth. It would seem likely that users of the system would typically be less interested in retrieving such short stories than those reported at more length. Therefore, the performance of the final system is probably better than is suggested by the recall scores. Furthermore, the evaluation demonstrates that the annotation system is very reliable, and therefore that searches performed using the search system would rarely return references to irrelevant media.

5.6 Summary of Applications

In this section we have discussed some examples of semantic web applications using Human Language Technology which have been developed specifically for real use in industry. While these are all research systems, they all demonstrate very clearly:

- the need for and importance of such applications in industry;
- the transition between research prototypes and real world applications;
- the actual use of such technologies in industry;
- the performance levels necessary to be of use in an industrial setting.

Some of the applications (such as Rich News and SWAN) are still in the process of development and improvement, but rather than being a drawback, this actually serves to emphasise the importance of benchmarking activities and testing applications in the real world, for it is often only through such methods (rather than laboratory testing under ideal conditions and with toy scenarios) that useful improvements can be made that will benefit end users to the maximum.

While the results of the best practices questionnaire show us the ideas and opinions about the Semantic Web of those involved in the field, the examples of real applications in use aim to show us in more practical terms which kind of practices are really useful. In particular, it shows that while these applications are not particularly mature – indeed, many of them area still ongoing further development and are still in the research phase – they are nevertheless useful to real users in industry as they stand. This emphasises the point that often tools and applications can be useful if only semi-automatic or if results are not perfect, because they enable users to save time and money in performing tasks which would previously would have been achieved manually, or with great difficulty by a human. Another particular point to note is that all these systems are based to some extent on the architecture on GATE, which was designed to be an open and flexible architecture for language processing. GATE has been used in many different ways, not just for information extraction (the application for which it is best known) but as the basis for many different tasks. It has also been designed to work with different languages, scenarios, applications, and domains with the maximum robustness and ease of adaptation. This, we believe, is a crucial point in expanding the prevalence of the Semantic Web.

A possible extension of this discussion could be in the future deliverable, to make a synthesis of implementers' opinions. This could be helpful in order to merge the global feedback, as provided by the previously mentioned questionnaire, and the concrete experience coming from tool development. Such experience can also be supplemented by a feedback from benchmarking evaluation. One of the difficulties of such a task is to identify a framework basis for the interview that would facilitate the synthesis of the feedback.

6 Discussion and future works

This deliverable is a first step toward a global document aiming at synthesising success stories and best practices of semantic Web technology. We insisted on the importance of going further than the simple technical aspects. Indeed, practices involve human factors and no technology can successfully emerge without taking into account the users. In our case the user is mainly a developer but this thinking can be extended to the final user.

This motivates us to start with an opinion poll that provides a feedback and shows us that, regarding concepts and practices of the Semantic Web, things are far from being obvious for all. This probably shows that a strategy progressing too fast toward best practice recommendations may not be a good solution. This does not mean that no recommendations should be made, but simply that at this stage of the semantic web evolution; best practice recommendations and education (i.e. teaching) should be considered together.

Even if a "best practices" guideline should remain the goal of our work, it is probably more reasonable to talk about frequent or consensus in practices than about best practices. This represents the second axis of our document that aims at extracting and suggesting practices from concrete and effective experience from several tools. As examples of such a recommendation, we could emphasise the preferences for simple architectures of services easily usable by non experts, the use of standard interfaces (e.g. for extraction from text). Of course, these are only basic illustrative examples but the scope of future work could be to survey those practices that are probably not "best practices" for all but that represent a trend or frequent practices.

In order to be constructive, these suggestions should be put in perspective from related initiatives. The survey of W3C activity and other related initiatives should continue and feed our next contribution.

7 References

[Cun02b] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002

[Pop02a] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff and M. Goranov, 2004. KIM -- Semantic Annotation Platform, *Journal of Natural Language Engineering*

[Maynard05b] D. Maynard, M. Yankova, A. Kourakis and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch, ESWC Workshop "End User Apects of the Semantic Web", Heraklion, Crete, 2005.

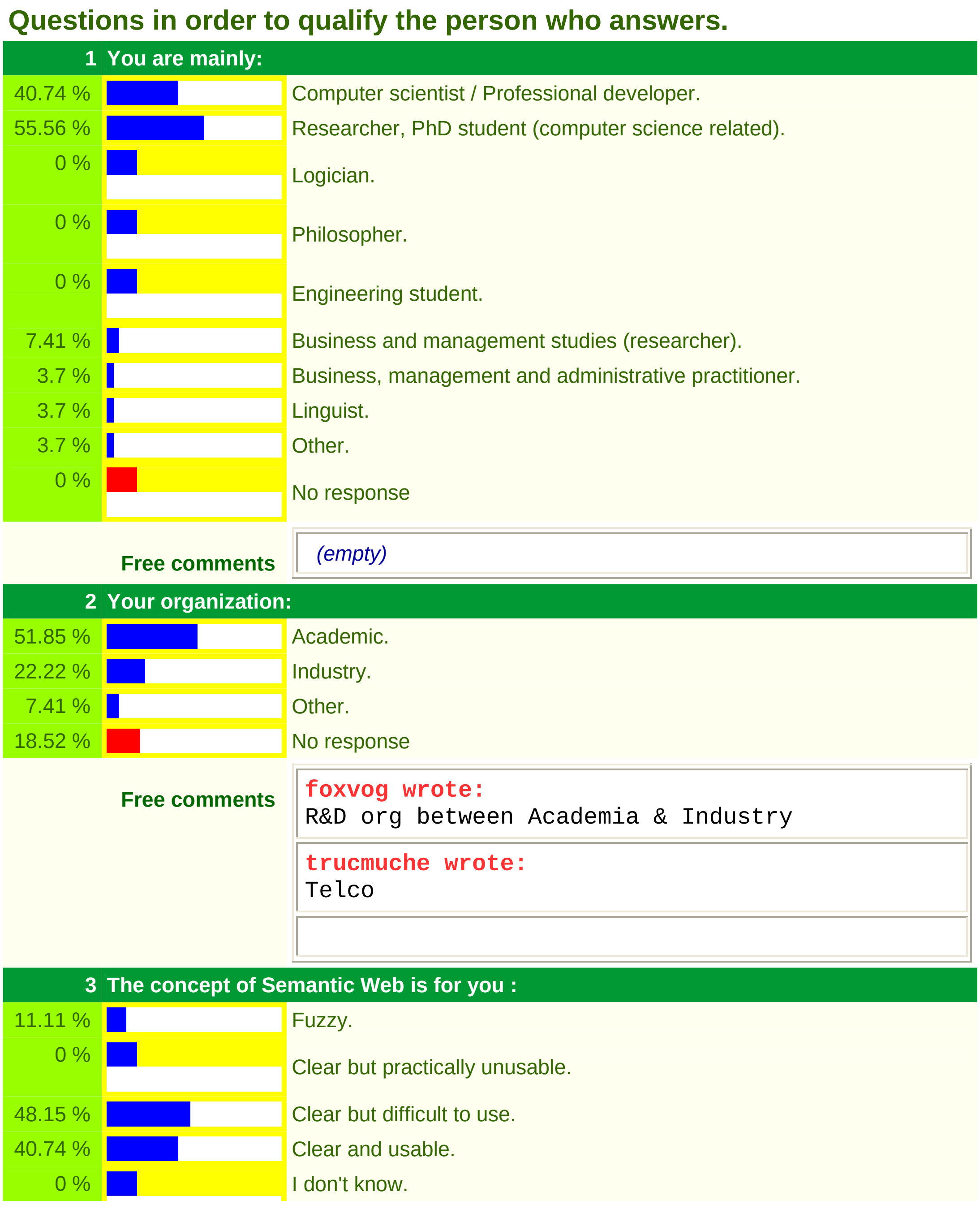
[May04b] D. Maynard, M. Yankova, N. Aswani, H. Cunningham. Automatic Creation and Monitoring of Semantic Metadata in a Dynamic Knowledge Portal. *Proceedings of the 11th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2004)*, Varna, Bulgaria, 2004.

[Dim04b] N. Dimitrova, J. Zimmerman, A. Janevski, L. Agnihotri, N. Haas, D. Li, R. Bolle, S. Velipasalar, T. McGee, L. Nikolovska. Media Personalisation and Augmentation Through Multimedia Processing and Information Extraction. In L. Ardissono and A. Kobsa and M. Maybury (eds.) *Personalised Digital Television.* Kluwer, pp. 201-233, 2004.

[Przybocki99] M. Przybocki, J. Fiscus, J. Garofolo and D. Pallett. 1998 {HUB-4} Information Extraction Evaluation, Proceedings of the DARPA Broadcast News Workshop, Herndon, VA, pp.13-18, 1999.

8 Annex

This section contains the raw results of 27 contributors at May 25, 2005.



Questions in order to qualify the person who answers.

1	You are mainly:
40.74 %	Computer scientist / Professional developer.
55.56 %	Researcher, PhD student (computer science related).
0 %	Logician.
0 %	Philosopher.
0 %	Engineering student.
7.41 %	Business and management studies (researcher).
3.7 %	Business, management and administrative practitioner.
3.7 %	Linguist.
3.7 %	Other.
0 %	No response

Free comments

(empty)

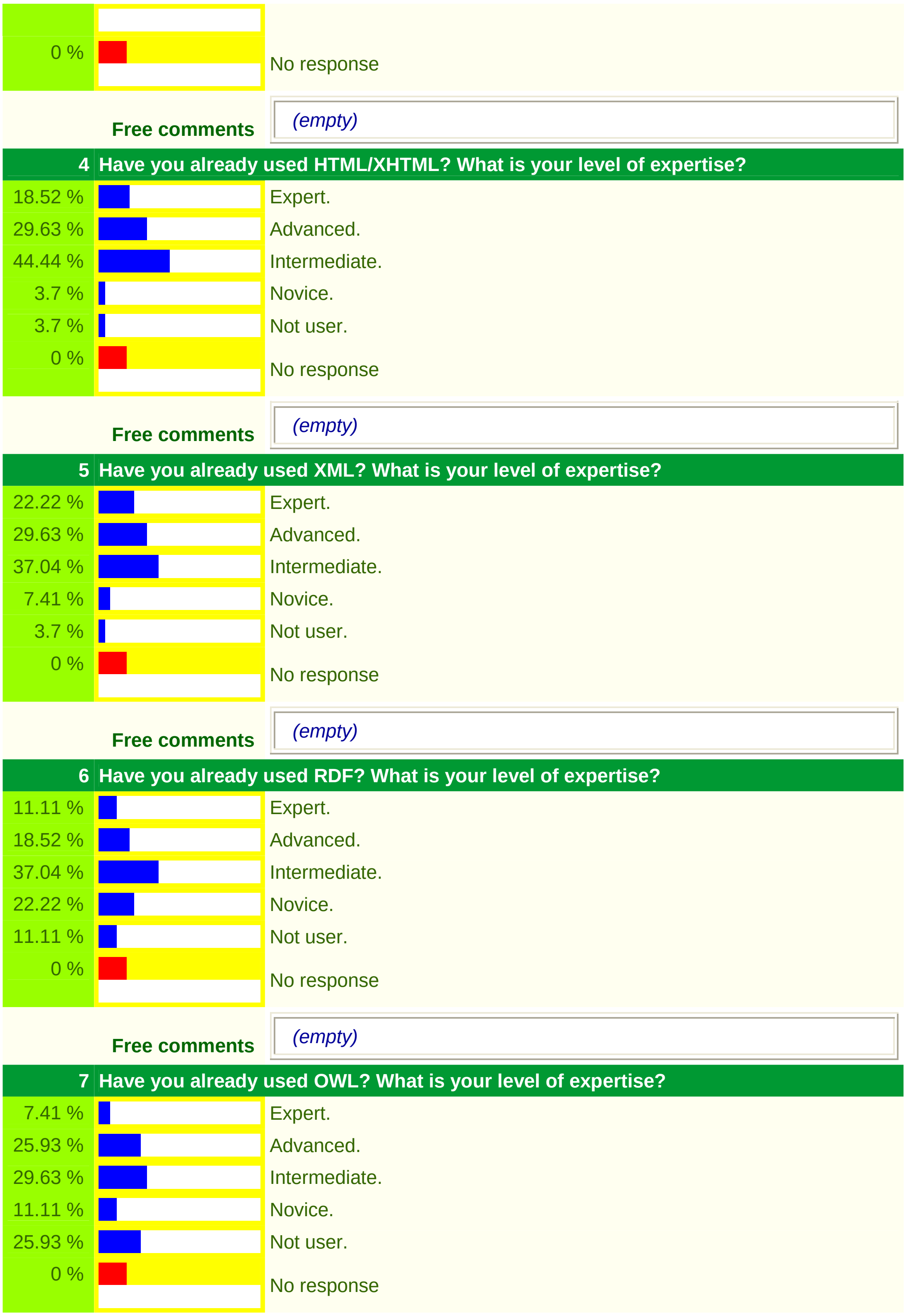
2	Your organization:
51.85 %	Academic.
22.22 %	Industry.
7.41 %	Other.
18.52 %	No response

Free comments

foxvog wrote:
R&D org between Academia & Industry

trucmuche wrote:
Telco

3	The concept of Semantic Web is for you :
11.11 %	Fuzzy.
0 %	Clear but practically unusable.
48.15 %	Clear but difficult to use.
40.74 %	Clear and usable.
0 %	I don't know.



0 %	🟥	No response
	Free comments	*(empty)*

4	Have you already used HTML/XHTML? What is your level of expertise?	
18.52 %	🟦	Expert.
29.63 %	🟦	Advanced.
44.44 %	🟦	Intermediate.
3.7 %	🟦	Novice.
3.7 %	🟦	Not user.
0 %	🟥	No response
	Free comments	*(empty)*

5	Have you already used XML? What is your level of expertise?	
22.22 %	🟦	Expert.
29.63 %	🟦	Advanced.
37.04 %	🟦	Intermediate.
7.41 %	🟦	Novice.
3.7 %	🟦	Not user.
0 %	🟥	No response
	Free comments	*(empty)*

6	Have you already used RDF? What is your level of expertise?	
11.11 %	🟦	Expert.
18.52 %	🟦	Advanced.
37.04 %	🟦	Intermediate.
22.22 %	🟦	Novice.
11.11 %	🟦	Not user.
0 %	🟥	No response
	Free comments	*(empty)*

7	Have you already used OWL? What is your level of expertise?	
7.41 %	🟦	Expert.
25.93 %	🟦	Advanced.
29.63 %	🟦	Intermediate.
11.11 %	🟦	Novice.
25.93 %	🟦	Not user.
0 %	🟥	No response

Free comments

(empty)

Goal of Best Pratices guidelines.

8		Usefulness of Best Practices Guide:
70.37 %	🟦	Necessary. There is a need for clarification in practices.
0 %	🟦	Not usable in practice. Developers should remain free since rules are mainly context dependent, and consequently difficult to generalize.
11.11 %	🟦	The usability of Best Practices guideliness is not clear.
18.52 %	🟥	No response

Free comments

franconi wrote:

```
Understanding (and teaching) semantics is more
important than best
practice
```

Frank.van.Harmelen@cs.vu.nl wrote:

```
W3C Semantic Web Best Practices working group
is doing important work
```

9		A Best Practices guideline should give:
62.96 %	🟦	High level advice (integration, interface, etc.). No technical advices because too difficult to generalize, only examples or successful stories adaptable by all developers.
44.44 %	🟦	Low level directives (standards remain open and can lead to anerror, a more strict guidance is necessary: a recommend format, structure, ...)
11.11 %	🟦	A technology based tutorial is enough.
0 %	🟦	I don't know.
7.41 %	🟥	No response

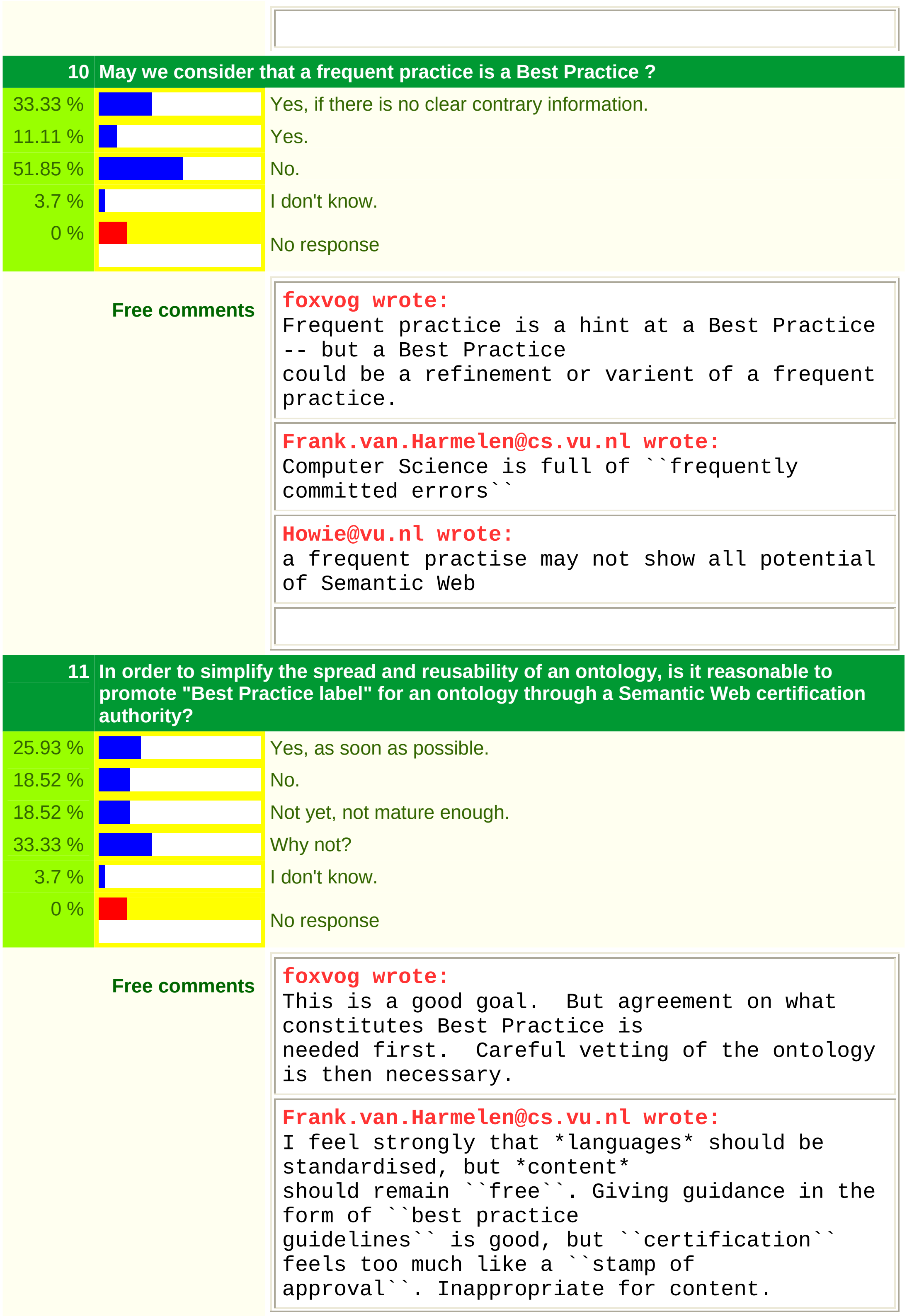
Free comments

foxvog wrote:

```
Both high-level advice and low level
directives
                        should be included
```

lyndon nixon wrote:

```
It could be extended by more specific cases,
when it is possible to
categorize Sem Web usage into different areas
(search, integration...)
We already maybe have a basis for this from
```



10	May we consider that a frequent practice is a Best Practice ?
33.33 %	Yes, if there is no clear contrary information.
11.11 %	Yes.
51.85 %	No.
3.7 %	I don't know.
0 %	No response

Free comments

```
foxvog wrote:
Frequent practice is a hint at a Best Practice
-- but a Best Practice
could be a refinement or varient of a frequent
practice.
```

```
Frank.van.Harmelen@cs.vu.nl wrote:
Computer Science is full of ``frequently
committed errors``
```

```
Howie@vu.nl wrote:
a frequent practise may not show all potential
of Semantic Web
```

11	In order to simplify the spread and reusability of an ontology, is it reasonable to promote "Best Practice label" for an ontology through a Semantic Web certification authority?
25.93 %	Yes, as soon as possible.
18.52 %	No.
18.52 %	Not yet, not mature enough.
33.33 %	Why not?
3.7 %	I don't know.
0 %	No response

Free comments

```
foxvog wrote:
This is a good goal.  But agreement on what
constitutes Best Practice is
needed first.  Careful vetting of the ontology
is then necessary.
```

```
Frank.van.Harmelen@cs.vu.nl wrote:
I feel strongly that *languages* should be
standardised, but *content*
should remain ``free``. Giving guidance in the
form of ``best practice
guidelines`` is good, but ``certification``
feels too much like a ``stamp of
approval``. Inappropriate for content.
```

Howie@vu.nl wrote:
Even industry needs some guideline, i.e. some expressive (and successful example) which can easily be adopted for their concrete use case.

Ontology and the real world.

12	Do you think that a philosopher should be involved in development of ontologies?
18.52 %	Mostly yes.
22.22 %	Mostly no (i.e too heavy to manage, only a computer scientist matter).
18.52 %	Depends on the size of the application.
11.11 %	Depends on the level of reusability of existing ontology.
48.15 %	Depends on the application.
0 %	No response

Free comments

Barry Norton wrote:
There are surely very engineering-oriented applications where there would be little input, subtle issues being dealt with only by importextension.

foxvog wrote:
Input from a philosopher can be useful if a trained ontological engineer is not available. However, philosophers may be too dogmatic and their rulings can be over-ruled by pragmatics.

franconi wrote:
It definitely helps for having a better ontology.

Howie@vu.nl wrote:
Delevelopment (and reuse) should strongly be application driven!

13	Do you think that a logician should be involved in development of ontologies?
22.22 %	Mostly yes.
7.41 %	Mostly no (i.e too heavy to manage, only a computer scientist matter).
14.81 %	Depends on the size of the application.

3.7 %	Depends on the level of reusability of existing ontology.
55.56 %	Depends on the application.
3.7 %	No response

Free comments

Barry Norton wrote:
Some ontologies will be little more than
taxonomies, with little
automated reasoning beyond subsumption, but
others will have more
advanced needs for automated reasoning etc.
that need to be anticipated
well.

foxvog wrote:
Someone with a working knowledge of formal
logic is needed.
Philosophers fall in this category.

franconi wrote:
Logicians are noit stricyly necessary, but may
help to teach the
languages involved.

Frank.van.Harmelen@cs.vu.nl wrote:
but only in an *advising* role, never in a
prescribing role

Howie@vu.nl wrote:
Delevelopment (and reuse) should strongly be
application driven!

lulu wrote:
What do you mean by logician ?

lyndon nixon wrote:
Some people use an ontology just like a simple
classification. Where it
is to be used by non-simple logical reasoning,
a logician becomes necessary.

wskw wrote:
Ontology design tool should point out logic
flaws (contradictions, etc.)

14	Do you think a linguist should be involved in development of ontologies?
33.33 %	Mostly yes.
3.7 %	Mostly no (i.e too heavy to manage, only a computer scientist matter).
14.81 %	Depends on the size of the application.
11.11 %	Depends on the level of reusability of existing ontology.
44.44 %	Depends on the application.
0 %	No response

Free comments

Barry Norton wrote:
Again, there will be engineering applications,
where there would be
little input, and knowledge acquisition
applications where this would be
essential.

foxvog wrote:
Linguists can be very useful, being aware of
ambiguities distinctions
that others may not notice. Useful for
designing equivalent names in
different contexts.

franconi wrote:
It definitely helps for having a better
ontology.

Frank.van.Harmelen@cs.vu.nl wrote:
again, only in an *advising* role, never in a
prescribing role

Howie@vu.nl wrote:
Delevelopment (and reuse) should strongly be
application driven!

lyndon nixon wrote:
Is the application NLP related or not?

15 Should we take into account the uncertainty or probabilities in concepts when developing ontology?

18.52 %	Strict yes.
48.15 %	Mostly yes.
18.52 %	Mostly no.
11.11 %	Strict no.
3.7 %	I don't know.
0 %	No response

Free comments

Barry Norton wrote:
(I`d rather have `somewhat yes`, but I guess
that`s a glass
half-fullempty distinction...)

foxvog wrote:
The fuzzyness of categories needs to be
recognized. But the assignment
of probabilities will usually not be needed.

franconi wrote:

It is a no if OWL or RDF is used, since these languages do not deal with uncertainty

Frank.van.Harmelen@cs.vu.nl wrote:
there are many domains where this is useful or even necessary. Too bad that RDF & OWL do not yet deal with Fuzzy concepts and relations

Howie@vu.nl wrote:
Only if there is a strong evidence in the application ;-)

lyndon nixon wrote:
I find in practice probabilistic logic is often an useful tool, if trying to model the ``real world``

wskw wrote:
Seems very important to me, but technology (even theory) is pretty immature in this area

Building ontology.

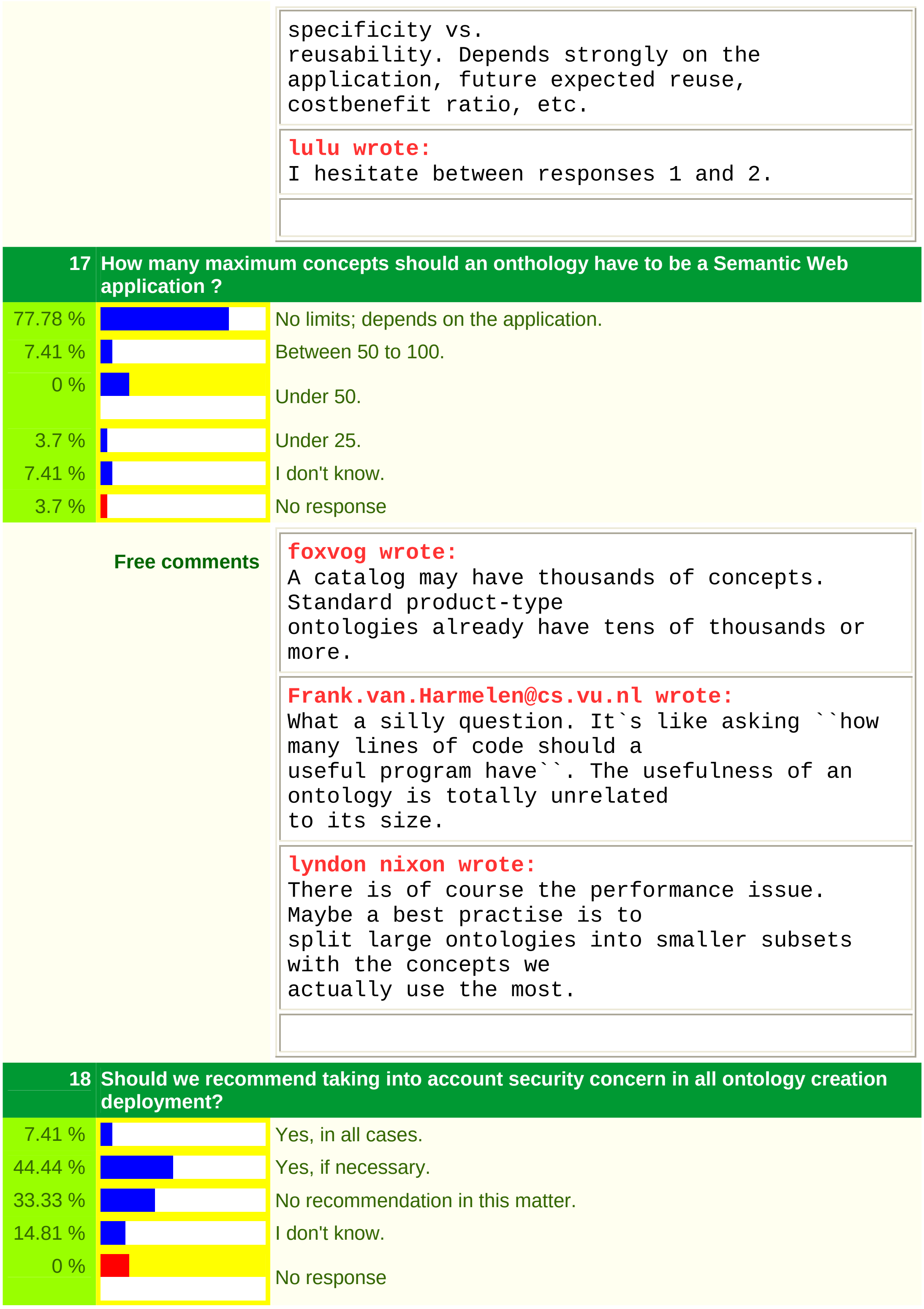
16	Should we recommend ...
55.56 %	Domain oriented ontology (best fit to the problem to solve).
29.63 %	General ontology, usable for a maximum of domains (best reusability ratio).
33.33 %	No rules in this matter.
0 %	I don't know.
3.7 %	No response

Free comments

Barry Norton wrote:
Should encourage reuse, but sensitive about preaching reusability (when not everyoneevery application is sufficiently developed, along these lines, to immediately consider this.)

foxvog wrote:
We should reccommend domain-oriented ontologies being linked to a general ontology. This enables easy incorporation of multiple ontologies in a system without having to download a universal ontology.

Frank.van.Harmelen@cs.vu.nl wrote:
this is a well=known trade-off in AICS: domain



specificity vs. reusability. Depends strongly on the application, future expected reuse, costbenefit ratio, etc.

lulu wrote:
I hesitate between responses 1 and 2.

17 How many maximum concepts should an onthology have to be a Semantic Web application ?

77.78 %	No limits; depends on the application.
7.41 %	Between 50 to 100.
0 %	Under 50.
3.7 %	Under 25.
7.41 %	I don't know.
3.7 %	No response

Free comments

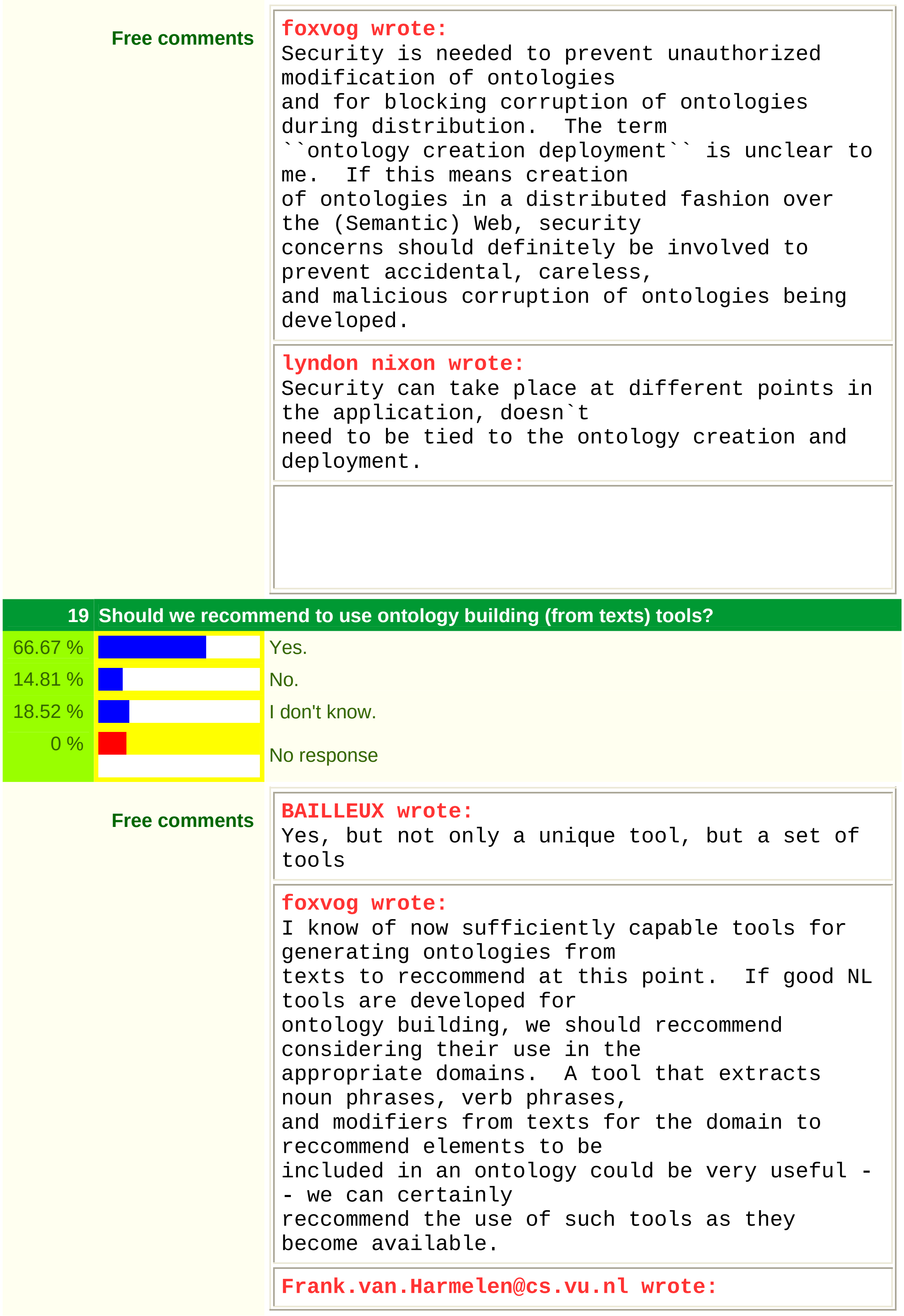
foxvog wrote:
A catalog may have thousands of concepts. Standard product-type ontologies already have tens of thousands or more.

Frank.van.Harmelen@cs.vu.nl wrote:
What a silly question. It`s like asking ``how many lines of code should a useful program have``. The usefulness of an ontology is totally unrelated to its size.

lyndon nixon wrote:
There is of course the performance issue. Maybe a best practise is to split large ontologies into smaller subsets with the concepts we actually use the most.

18 Should we recommend taking into account security concern in all ontology creation deployment?

7.41 %	Yes, in all cases.
44.44 %	Yes, if necessary.
33.33 %	No recommendation in this matter.
14.81 %	I don't know.
0 %	No response



Free comments

foxvog wrote:
Security is needed to prevent unauthorized modification of ontologies and for blocking corruption of ontologies during distribution. The term ``ontology creation deployment`` is unclear to me. If this means creation of ontologies in a distributed fashion over the (Semantic) Web, security concerns should definitely be involved to prevent accidental, careless, and malicious corruption of ontologies being developed.

lyndon nixon wrote:
Security can take place at different points in the application, doesn`t need to be tied to the ontology creation and deployment.

19 Should we recommend to use ontology building (from texts) tools?

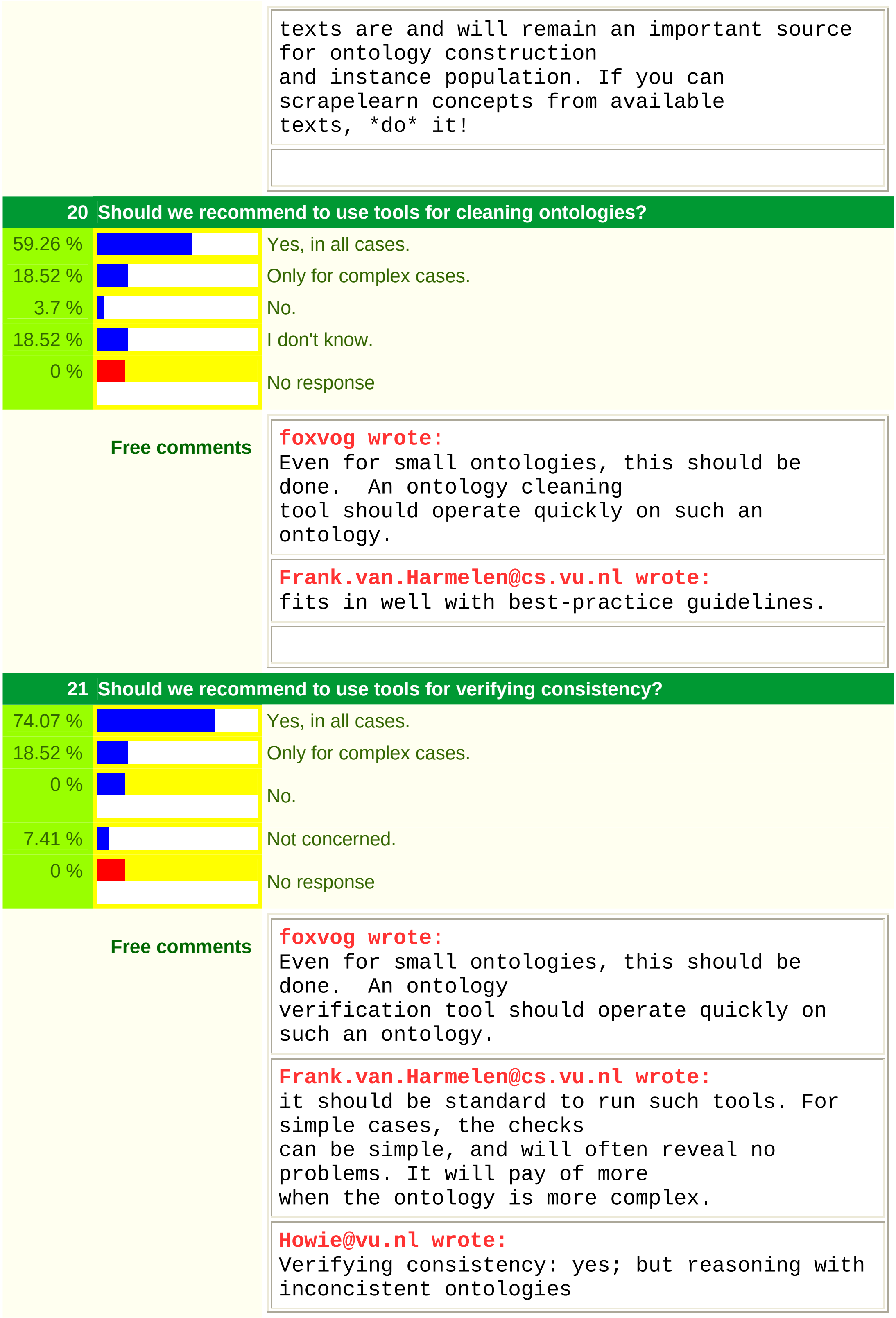
66.67 %	Yes.
14.81 %	No.
18.52 %	I don't know.
0 %	No response

Free comments

BAILLEUX wrote:
Yes, but not only a unique tool, but a set of tools

foxvog wrote:
I know of now sufficiently capable tools for generating ontologies from texts to reccommend at this point. If good NL tools are developed for ontology building, we should reccommend considering their use in the appropriate domains. A tool that extracts noun phrases, verb phrases, and modifiers from texts for the domain to reccommend elements to be included in an ontology could be very useful -- we can certainly reccommend the use of such tools as they become available.

Frank.van.Harmelen@cs.vu.nl wrote:



texts are and will remain an important source
for ontology construction
and instance population. If you can
scrapelearn concepts from available
texts, *do* it!

20 Should we recommend to use tools for cleaning ontologies?

59.26 %	Yes, in all cases.
18.52 %	Only for complex cases.
3.7 %	No.
18.52 %	I don't know.
0 %	No response

Free comments

foxvog wrote:
Even for small ontologies, this should be
done. An ontology cleaning
tool should operate quickly on such an
ontology.

Frank.van.Harmelen@cs.vu.nl wrote:
fits in well with best-practice guidelines.

21 Should we recommend to use tools for verifying consistency?

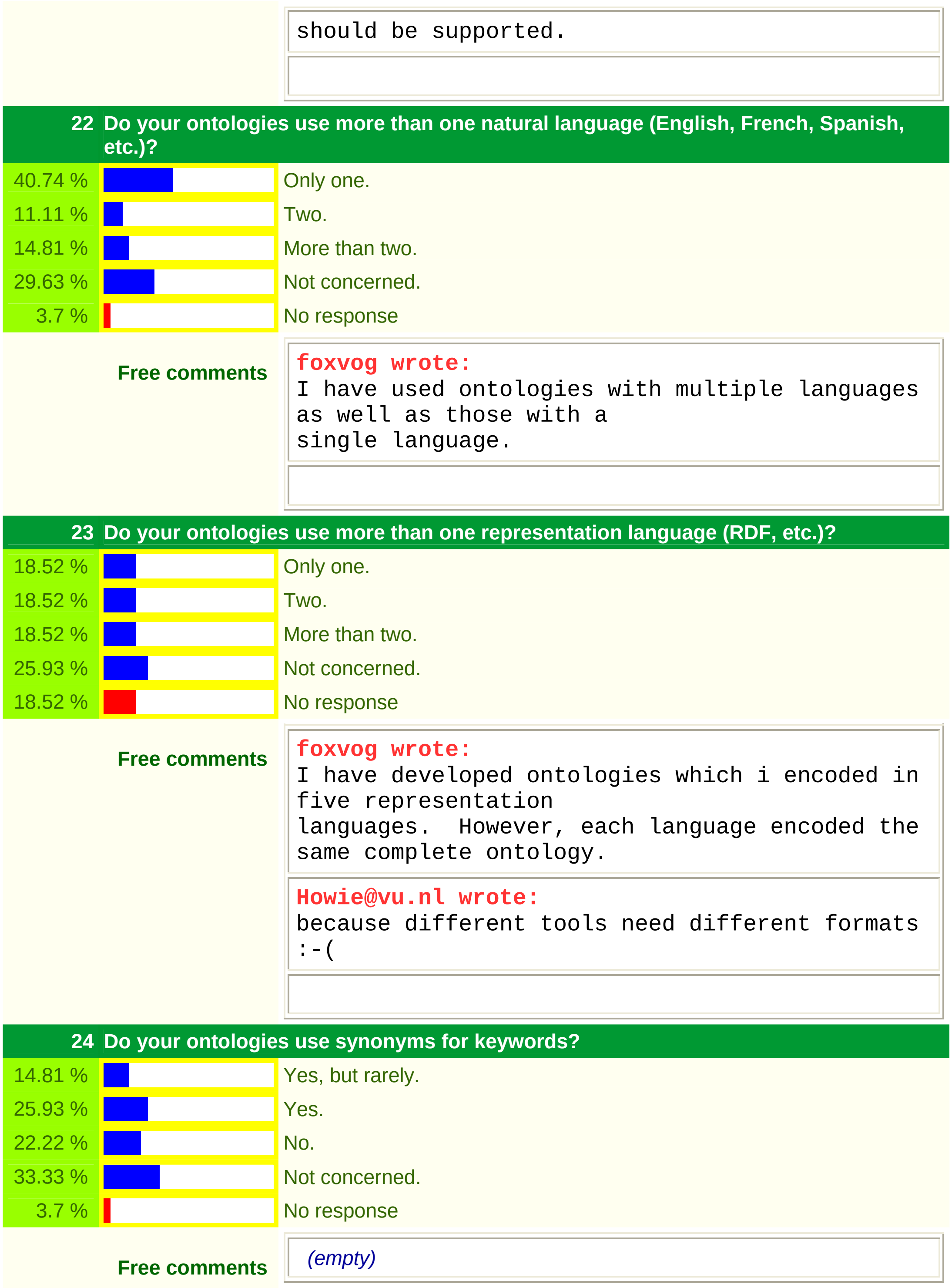
74.07 %	Yes, in all cases.
18.52 %	Only for complex cases.
0 %	No.
7.41 %	Not concerned.
0 %	No response

Free comments

foxvog wrote:
Even for small ontologies, this should be
done. An ontology
verification tool should operate quickly on
such an ontology.

Frank.van.Harmelen@cs.vu.nl wrote:
it should be standard to run such tools. For
simple cases, the checks
can be simple, and will often reveal no
problems. It will pay of more
when the ontology is more complex.

Howie@vu.nl wrote:
Verifying consistency: yes; but reasoning with
inconcistent ontologies



should be supported.

22 Do your ontologies use more than one natural language (English, French, Spanish, etc.)?

40.74 %	Only one.
11.11 %	Two.
14.81 %	More than two.
29.63 %	Not concerned.
3.7 %	No response

Free comments

foxvog wrote:
I have used ontologies with multiple languages as well as those with a single language.

23 Do your ontologies use more than one representation language (RDF, etc.)?

18.52 %	Only one.
18.52 %	Two.
18.52 %	More than two.
25.93 %	Not concerned.
18.52 %	No response

Free comments

foxvog wrote:
I have developed ontologies which i encoded in five representation languages. However, each language encoded the same complete ontology.

Howie@vu.nl wrote:
because different tools need different formats :-(

24 Do your ontologies use synonyms for keywords?

14.81 %	Yes, but rarely.
25.93 %	Yes.
22.22 %	No.
33.33 %	Not concerned.
3.7 %	No response

Free comments *(empty)*

Availability of ontology.

25	Regarding ontology time to live:
85.19 %	Ontologies are supposed to be persitent for a long period (can be used by several generations of application)
3.7 %	Ontology time to live is limited to the duration of a dedicated application (involve low reusability)
7.41 %	I don't know.
3.7 %	No response

Free comments

```
Barry Norton wrote:
(No `depends on application` here?  That would
have been my choice,
although with aims towards the long-term...)
```

```
foxvog wrote:
Both are possible.  We should strive for
reuse.  Modularization will
help ontology reuse.  We should reccommend
modularized ontologies.
```

```
Frank.van.Harmelen@cs.vu.nl wrote:
if the Semantic Web is to live at all,
ontologies shoud be reused across
applications, and across multiple generations
of the same application.
```

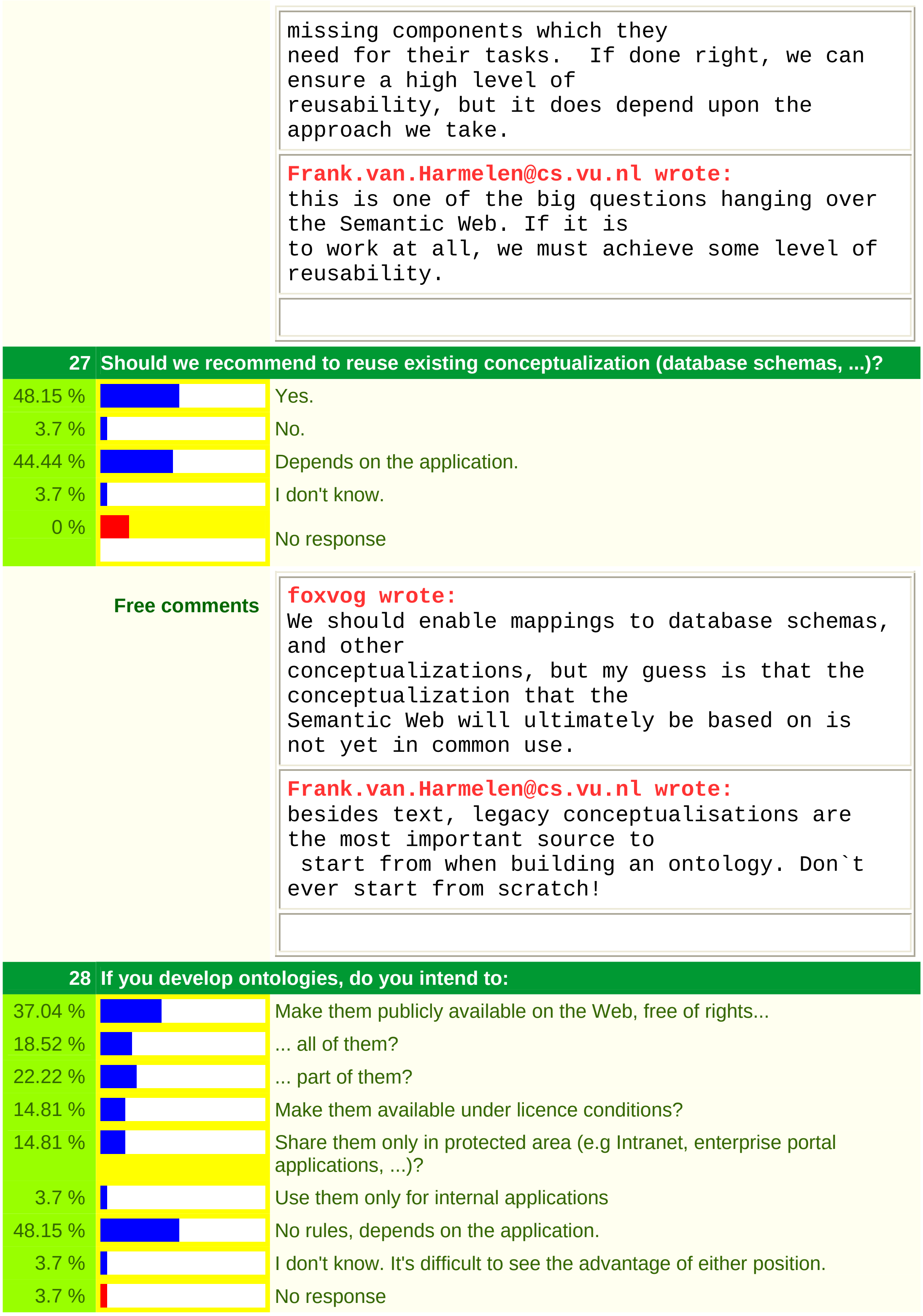
```
RobertTolksdorf wrote:
This implies the need for practices to
maintain ontologies.
```

```
wskw wrote:
Depends on application.
```

26	Regarding the reusability of an ontology, do you think that:
29.63 %	In practice the level of reusability will be very low because many developers create ontologies first of all for their own objective.
18.52 %	There is no doubt that Semantic Web technologies will draw to a high level of reusability.
48.15 %	We hope to have a fair level of reusability but it is not clear.
0 %	I don't know.
18.52 %	No response

Free comments

```
foxvog wrote:
Popularizing well modularized ontologies will
draw a higher level of
reusability.  If made available, developers
will select modules for
reuse instead of spending time re-inventing
the wheel, IFF they are able
to add their own modules to include the
```



missing components which they
need for their tasks. If done right, we can
ensure a high level of
reusability, but it does depend upon the
approach we take.

Frank.van.Harmelen@cs.vu.nl wrote:
this is one of the big questions hanging over
the Semantic Web. If it is
to work at all, we must achieve some level of
reusability.

27 Should we recommend to reuse existing conceptualization (database schemas, ...)?

48.15 %	Yes.
3.7 %	No.
44.44 %	Depends on the application.
3.7 %	I don't know.
0 %	No response

Free comments

foxvog wrote:
We should enable mappings to database schemas,
and other
conceptualizations, but my guess is that the
conceptualization that the
Semantic Web will ultimately be based on is
not yet in common use.

Frank.van.Harmelen@cs.vu.nl wrote:
besides text, legacy conceptualisations are
the most important source to
start from when building an ontology. Don`t
ever start from scratch!

28 If you develop ontologies, do you intend to:

37.04 %	Make them publicly available on the Web, free of rights...
18.52 %	... all of them?
22.22 %	... part of them?
14.81 %	Make them available under licence conditions?
14.81 %	Share them only in protected area (e.g Intranet, enterprise portal applications, ...)?
3.7 %	Use them only for internal applications
48.15 %	No rules, depends on the application.
3.7 %	I don't know. It's difficult to see the advantage of either position.
3.7 %	No response

Free comments

Barry Norton wrote:
Intention is free (as in beer) publication,
reality is sometimes not,
depending on application.

foxvog wrote:
The general parts of ontologies, i would want
to make publically
available. Narrow domain ontologies would be
only provided under
license conditions according to my
organization`s current policies.
I would prefer a licensing scheme in which
free use and expansion of the
ontologies is permitted, but vetting of
modification of existing
components is required so that a babble of
inconsistent versions does
not evolve.

29 Should we recommend RDF formalism only for ontology already available in other knowledge representation formalism (trees, ...) i.e translation of representation?

33.33 %	Yes, if possible.
11.11 %	Yes, a dedicated effort should be lunched in order to deliver RDF version of other ontology.
22.22 %	No. (e.g: too complex to manage; better to rewrite; inadequate descriptive capacities, ...
25.93 %	I don't know.
7.41 %	No response

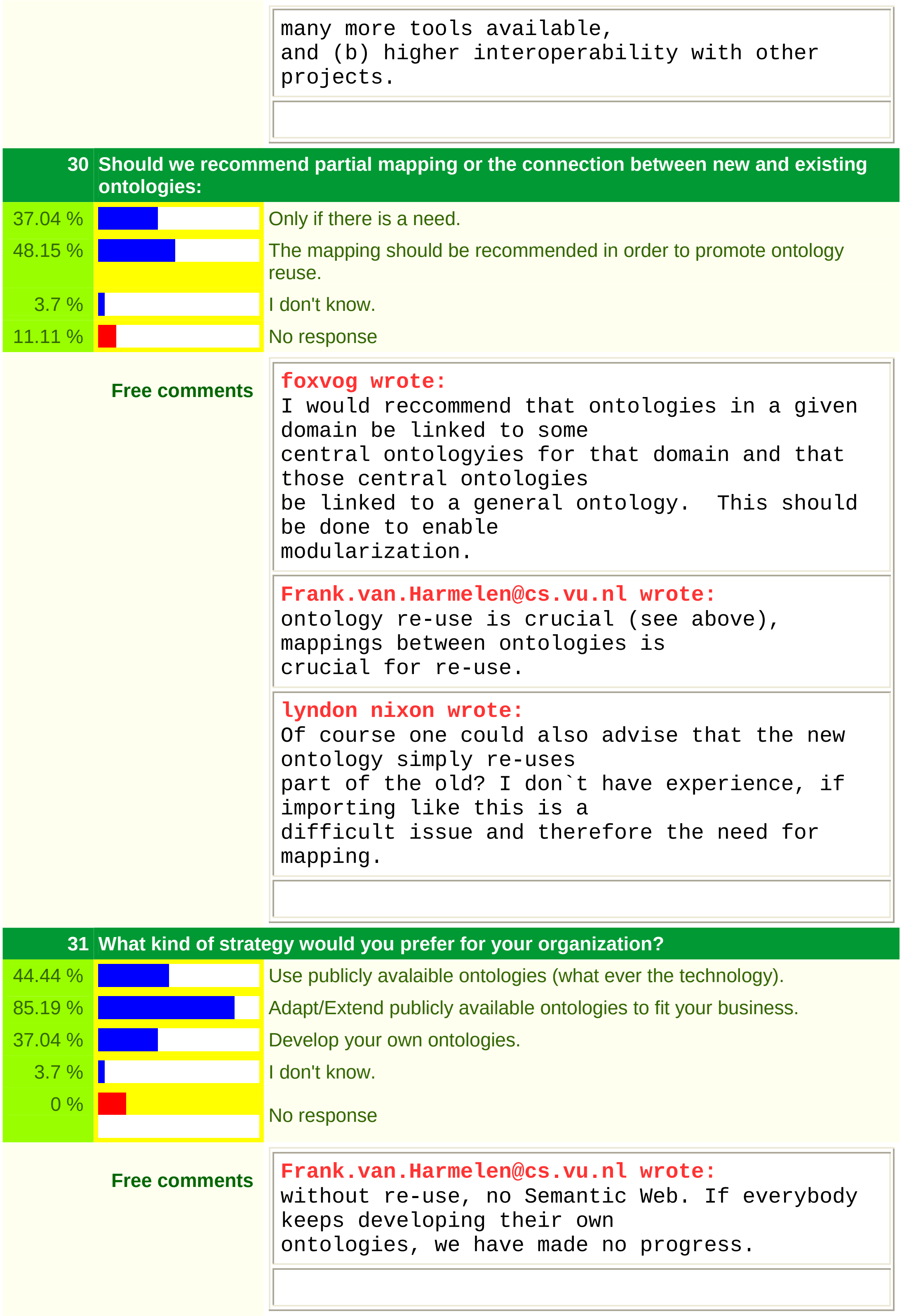
Free comments

Anna wrote:
the question is unclear

Barry Norton wrote:
(lunched?!?)

foxvog wrote:
RDF triples are quite restrictive. We should
envisage an evolution
beyond this formalism. As an interim step,
mapping other ontolgies to
RDF can be useful -- we should suggest that
those who want reuse of
other ontologies consider creating such
mappings, but not reccommend
that they do so no matter what.

Frank.van.Harmelen@cs.vu.nl wrote:
I`ve seen many projects that benefitted
enormously from exporting their
ontology to RDFOWL, and getting thereby (a)



```
many more tools available,
and (b) higher interoperability with other
projects.
```

30 Should we recommend partial mapping or the connection between new and existing ontologies:

37.04 %	Only if there is a need.
48.15 %	The mapping should be recommended in order to promote ontology reuse.
3.7 %	I don't know.
11.11 %	No response

Free comments

```
foxvog wrote:
I would reccommend that ontologies in a given
domain be linked to some
central ontologyies for that domain and that
those central ontologies
be linked to a general ontology.  This should
be done to enable
modularization.
```

```
Frank.van.Harmelen@cs.vu.nl wrote:
ontology re-use is crucial (see above),
mappings between ontologies is
crucial for re-use.
```

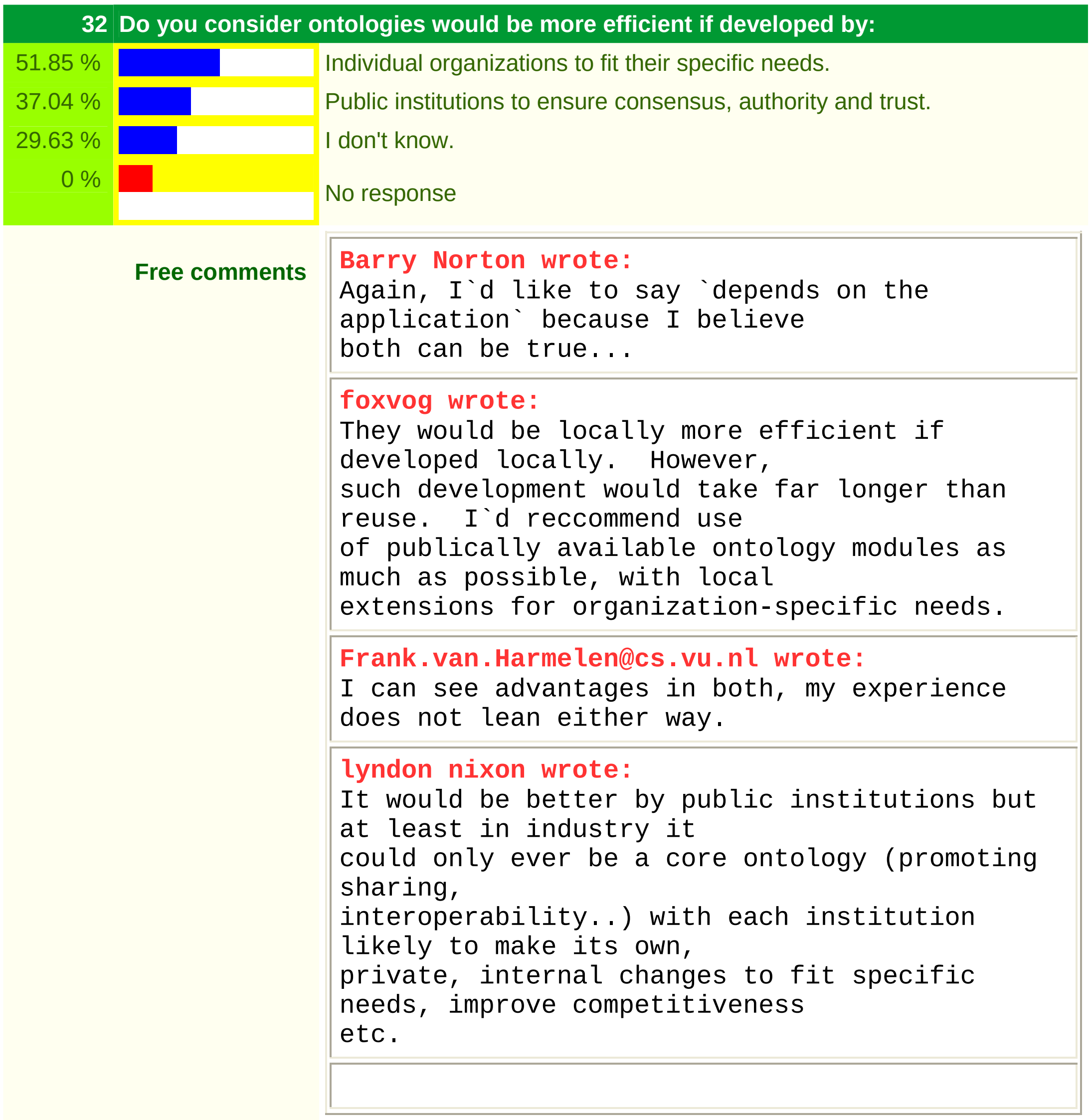
```
lyndon nixon wrote:
Of course one could also advise that the new
ontology simply re-uses
part of the old? I don`t have experience, if
importing like this is a
difficult issue and therefore the need for
mapping.
```

31 What kind of strategy would you prefer for your organization?

44.44 %	Use publicly avalaible ontologies (what ever the technology).
85.19 %	Adapt/Extend publicly available ontologies to fit your business.
37.04 %	Develop your own ontologies.
3.7 %	I don't know.
0 %	No response

Free comments

```
Frank.van.Harmelen@cs.vu.nl wrote:
without re-use, no Semantic Web. If everybody
keeps developing their own
ontologies, we have made no progress.
```



32	Do you consider ontologies would be more efficient if developed by:
51.85 %	Individual organizations to fit their specific needs.
37.04 %	Public institutions to ensure consensus, authority and trust.
29.63 %	I don't know.
0 %	No response

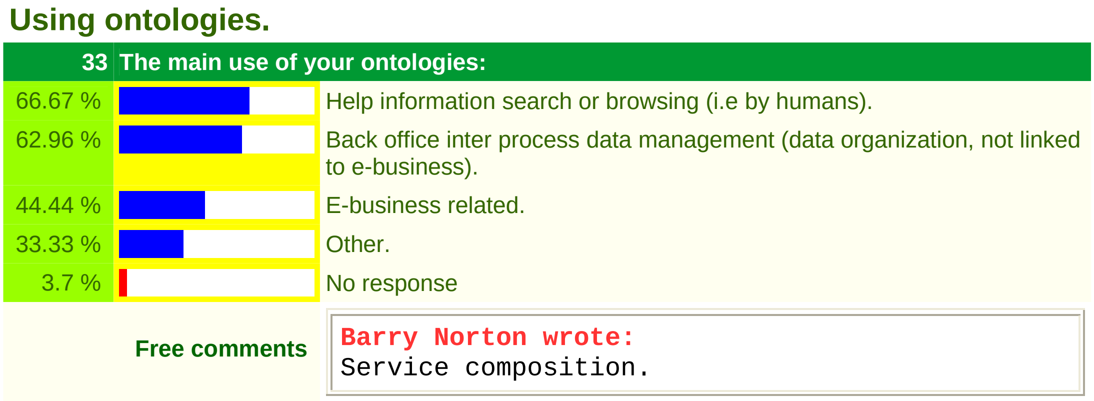
Free comments

```
Barry Norton wrote:
Again, I`d like to say `depends on the
application` because I believe
both can be true...
```

```
foxvog wrote:
They would be locally more efficient if
developed locally.  However,
such development would take far longer than
reuse.  I`d reccommend use
of publically available ontology modules as
much as possible, with local
extensions for organization-specific needs.
```

```
Frank.van.Harmelen@cs.vu.nl wrote:
I can see advantages in both, my experience
does not lean either way.
```

```
lyndon nixon wrote:
It would be better by public institutions but
at least in industry it
could only ever be a core ontology (promoting
sharing,
interoperability..) with each institution
likely to make its own,
private, internal changes to fit specific
needs, improve competitiveness
etc.
```

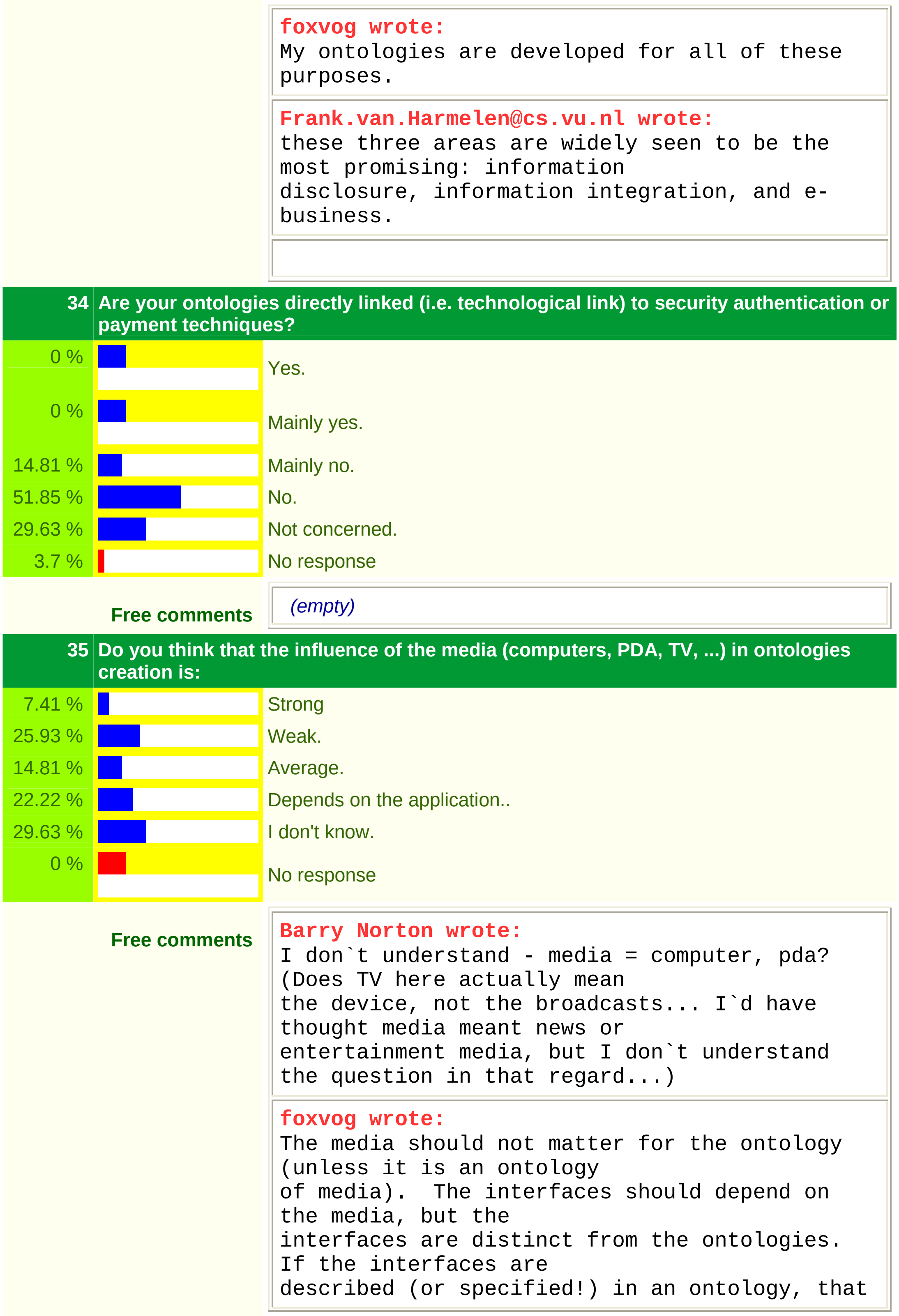


Using ontologies.

33	The main use of your ontologies:
66.67 %	Help information search or browsing (i.e by humans).
62.96 %	Back office inter process data management (data organization, not linked to e-business).
44.44 %	E-business related.
33.33 %	Other.
3.7 %	No response

Free comments

```
Barry Norton wrote:
Service composition.
```



foxvog wrote:
My ontologies are developed for all of these purposes.

Frank.van.Harmelen@cs.vu.nl wrote:
these three areas are widely seen to be the most promising: information disclosure, information integration, and e-business.

34 Are your ontologies directly linked (i.e. technological link) to security authentication or payment techniques?

0 %	Yes.
0 %	Mainly yes.
14.81 %	Mainly no.
51.85 %	No.
29.63 %	Not concerned.
3.7 %	No response

Free comments

(empty)

35 Do you think that the influence of the media (computers, PDA, TV, ...) in ontologies creation is:

7.41 %	Strong
25.93 %	Weak.
14.81 %	Average.
22.22 %	Depends on the application..
29.63 %	I don't know.
0 %	No response

Free comments

Barry Norton wrote:
I don`t understand - media = computer, pda? (Does TV here actually mean the device, not the broadcasts... I`d have thought media meant news or entertainment media, but I don`t understand the question in that regard...)

foxvog wrote:
The media should not matter for the ontology (unless it is an ontology of media). The interfaces should depend on the media, but the interfaces are distinct from the ontologies. If the interfaces are described (or specified!) in an ontology, that

KB certainly is
influenced by the data (media) which it is
describing.

Technical concerns.

36	In order to use the semantics of data, we need (i.e. it's enough):
11.11 %	Proprietary format adapted to the application.
3.7 %	XHTML and Meta Tag are enough.
3.7 %	XML basic.
37.04 %	RDF.
70.37 %	RDF and OWL.
11.11 %	Other.
7.41 %	No response

Free comments

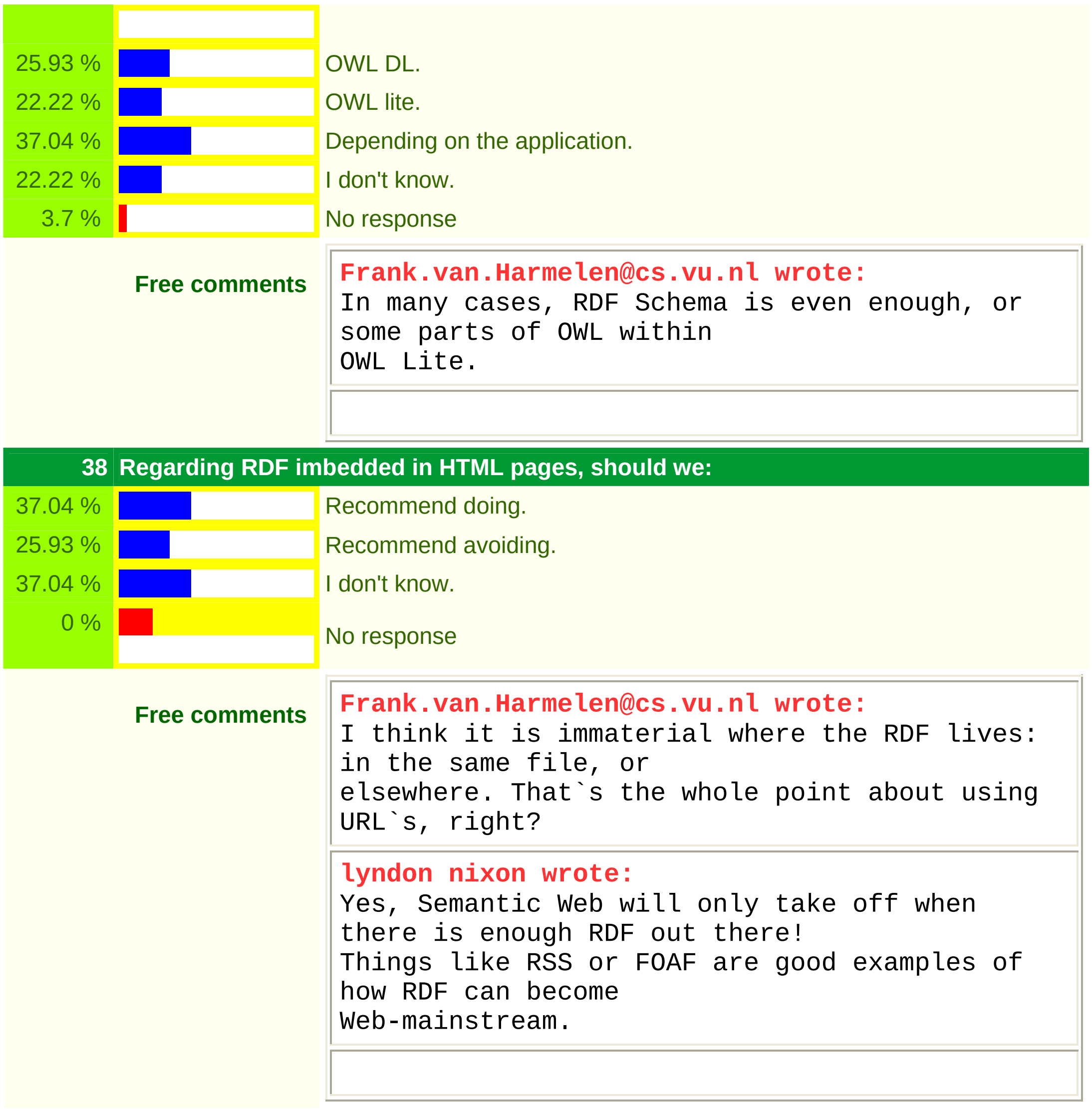
foxvog wrote:
RDF and OWL are sufficient, but other
formalisms (e.g. WSML) can be
sufficient as well. No single formalism is
needed.

franconi wrote:
Logic based ontology languages

Frank.van.Harmelen@cs.vu.nl wrote:
clearly proprietary formats won`t work (no
interoperability); XHTML and
XML won`t work (no semantics); RDF (including
of course RDF Schema) will
work in many cases (also in my experience),
OWL will be needed for more
complex cases.

lyndon nixon wrote:
how do you mean ``use``? Its clear that XML
basic can be enough to make
``intelligent`` tools (see how people use the
XML from Amazon or Google).
However in real world data processing, even
RDF is typically not enough.
So I answer RDF & OWL, but I find also in many
scenarios that this is
still not enough, i.e. we need also the Rules
layer!

37	Do you prefer the use of:
0 %	OWL full.



25.93 %	OWL DL.
22.22 %	OWL lite.
37.04 %	Depending on the application.
22.22 %	I don't know.
3.7 %	No response

Free comments

Frank.van.Harmelen@cs.vu.nl wrote:
In many cases, RDF Schema is even enough, or
some parts of OWL within
OWL Lite.

38 Regarding RDF imbedded in HTML pages, should we:

37.04 %	Recommend doing.
25.93 %	Recommend avoiding.
37.04 %	I don't know.
0 %	No response

Free comments

Frank.van.Harmelen@cs.vu.nl wrote:
I think it is immaterial where the RDF lives:
in the same file, or
elsewhere. That`s the whole point about using
URL`s, right?

lyndon nixon wrote:
Yes, Semantic Web will only take off when
there is enough RDF out there!
Things like RSS or FOAF are good examples of
how RDF can become
Web-mainstream.