INFO116 Group Assignment: Making Sense of History

You will be working in groups of 4 within your lab classes. You will have to manage the project in your groups, working in your own time, but you will also have some lab sessions where you can work on the assignment. In the first project management lab session the instructors will give an introduction to project management. While this course isn't about project management, you should keep in mind that the more successful groups will be the ones who can manage the tasks between the individual members properly. At the end of the projects, EACH MEMBER MUST be clearly accountable for their contribution.

Your overall task is to enhance a website with semantics, and at the same time to learn about history.
The web site is provided by the British Library (BL) Newspaper Archives.

https://www.britishnewspaperarchive.co.uk/

Unfortunately this is a paid service, so you cannot access it freely. (If you make an account, you can have three free searches.) I signed up for the purposes of this assignment and downloaded some web pages which I will provide to you for markup.
**** UPDATE *****
I managed to obtain a limited number of free logins for this assignment from the British Library. Details will follow.

We want to enhance the website with semantics for the following reasons:

- Make the site more friendly for search engines with better schema.org markup.
- Make the site Facebook friendly with OGP Basic Metadata.
- Include your own ontology to make the information about the page explicit and machine readable. This is the meat of your assignment. You will develop an ontology to capture the semantics of the articles.

Tasks

These are the specific tasks you will need to perform in this assignment:

1. Create an ontology of concepts for the historical articles, using Protege.
2. Write some SPARQL queries to show how an application could use the Ontology (and therefore the semantic data embedded in the web pages). You do not have to write any applications, only write down what queries they might use, and what answers you would get, and why those answers are useful.
3. Annotate ONE OF the web pages with semantic data, including your ontology and schema.org. The annotation will be both JSON-LD and RDFa Lite. In addition, add some OGP Basic Metadata for Facebook.
4. Write a project report. The report should be no longer than 10 pages, including figures, example markup, queries, etc.

Procedure

You should make an ontology with Protege, and create classes and properties that capture the domain of historical events. The ontology will serve two purposes. First, it is a model which allows you to answer interesting SPARQL queries. Second, you will annotate specific instances in the BL web pages with the relevant ontology classes. These will be inserted into the web page as RDFa Lite and JSON-LD.

The ontology should be able to represent details of the events depicted in the articles provided to you. In addition, it should be able to represent the events in the document “The Most Important Events of the Century From the Viewpoint of the People”. You should create competency questions to help you construct the ontology. For example:

- When did WWII start?
- Who were the major players in WWII?
- What is Civil Rights Movement?
- Who was the first person to land on the Moon?

Clearly you will have to limit the size of your ontology, and it won’t be an easy task. Hopefully it will be fun.

Here are some tips to help you.

Use existing ontologies!! You should use one of the powerful features available to Ontology engineers, the *import* imperative. You should import an existing ontology if you plan to use most of its classes and properties in your own ontology. However, importing is not always the best thing to do, especially if you only want to use a few classes from a very large existing ontology (importing an ontology brings ALL its existing classes into your own).

Sometimes you might just want to copy some of the classes into your ontology. The following stackoverflow answer explains how to do this.

https://stackoverflow.com/questions/44205661/how-to-import-specific-classes-and-object-properties-from-an-ontology-in-protege

You should read about ontology imports. It is a fairly straightforward idea, though in practice it can become complicated, especially when the name of an ontology does not correspond to its web location, etc.
Here is an introduction to imports in Protege:

https://protegewiki.stanford.edu/wiki/Importing_Ontologies_in_P41

In order to use the schema.org concepts, you can use the RDF representation of schema.org from the following URL:

https://schema.org/docs/developers.html#formats

There are many other resources to help you. I found a large number of free newspapers archives through a web search. Wikipedia has an excellent article with lots of links. Most of these sites have lots of schema.org and ogp markup.

https://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives

Another interesting resource is our own Marcus library site. This is a little different in that it catalogues historical artifacts, but they have an ontology of historical events. You have already seen a bit of this site in the seminar groups. The SPARQL endpoint is also a resource they provide to help study the ontology and contents:

http://sparql.ub.uib.no/

(Ignore /wab … that is data for the Wittgenstein archives).

Finally there is the mother of all historical ontologies, the cidoc-crm. This is a rather large and complex ontology, but very useful for getting ideas about modeling historical events, time periods, and so on.

http://www.cidoc-crm.org/

The functional overview might be helpful:

http://www.cidoc-crm.org/functional-units

While you are building your ontology you should also be building a collection of SPARQL queries that can answer the interesting competency questions using the ontology.

Finally note that the ontology is a little strange in that there are not so many individual instances. The individuals will be the articles, or the OCR fragments within the articles.

Annotation using OGP, RDFa Lite and JSON-LD

You should add the relevant information to the web pages, using both RDFa Lite and JSON-LD, as well as OGP (just some basic metadata). When you add the ontology data, only add the relevant concepts and properties, not the whole ontology!!

The web pages are a bit tricky, and you will need some ingenuity to find the text.

You should test the markup with the google testing tool, and the linter. Sometimes one works better than the other, so always try both!

https://search.google.com/structured-data/testing-tool#

http://linter.structured-data.org/

Try also the tool from http://rdf-translator.appspot.com/
Though I have not had much luck with it on these documents. Extra credits for HTML wizzards who get it working!

Deliverables

This is assignment requires a fair bit of work divided among the group members, but it should not be that difficult. A good pass can be obtained with appropriate and technically correct schema.org and OGP markup, a sensible basic ontology, adding some data to the web site with RDFa Lite and JSON-LD, and getting a few basic SPARQL queries working. The highest marks will go to groups with creative uses of the markup, and more extensive and complete use cases covered, and good justifications for their particular use of markup.

The deliverables will be

1. A written report (explained below)
2. The annotated HTML page you have chosen.
3. The ontology as an OWL file
4. All SPARQL queries and result sets. You should have a minimum of 5 non trivial queries in a text file.

The report should present an overview of what has been achieved. What work did each group member contribute? Why was the ontology constructed in the way it was? What kinds of questions can be answered by the ontology (the competency questions you used)? What can the web site do with the added semantics (e.g. third party applications)? The report should also include examples of the expected rich snippet from the markup (from the rich snippet tool). The report should be no longer than 10 pages, including figures, example markup, queries, etc.

The deliverables should be bundled into a zip file and submitted. You will get information about how to identify the documents before you submit. Typically you do not put your names on the report, but only the candidate numbers of every person in the group.

Submission deadline is November 27 at 14:00.