tarballs from wikimedia.org, tools.powerset.com

Content Repository implemented on top of two Amazon services: S3 and SDS
SDS is a columned and indexed distributed db, storing small key/value pairs per row
S3 is a storage service with retrieval via ID, limited metadata via unindexed key/value pairs

Content Repository:
doc CONTENT fields

Content Repository:
doc TEXT fields

Content Repository:
Raw source data, e.g. wikipedia xml dump

**Content Uploader**
Divides into individual articles, stores each in CR
(Web crawler would go here)

**Text Extraction (CTE)**
Hadoop job runs text_extract script on each article

**Resources**

**Permanent Storage**

**Data Description**

**Permanent Storage**

**Data Description**

**Programs and Actions**

**Sentence Breaking**
by XLE/C-FSM or just FST

**Named Entity Tagging**
by KR Lexicon-as-FST-regex

**Tokenizer+Morphology**
by XLE/C-FSM
(Transducer derived from Inxight)

**Parsing**
by the XLE

**Semrepping**
by the XLE transfer system
(implemented in Sicstus Prolog)

**Create index slivers**
by Quantum OR Chewie

**Index Repository**

**Index Slivers**

END CoralScript Pipeline processing phase

START CoralScript Pipeline processing phase

Below this line is is the CoralInterpreter CoralScript Pipeline: per-silver, per-article processing on a single local machine.

Data has hierarchical organization:  Project => (Slices) => Workunits => Slivers => Articles => Sentences
One Workunit = One Hadoop job. The upper levels of hierarchy are necessitated by limitations of Hadoop.
The Hadoop MapTask consists of sequentially executing all pipeline steps on a single Silver.
The sequence of pipeline steps and their parameters are defined by a CoralScript.

Bulk of processing occurs here: mostly Parsing, and secondarily Semrepping.

The Index!

wikipedia:
~300-400 GB

END CoralScript Pipeline processing phase

Brendan O’Connor, 7/29/2007