Importing Wikipedia in Plone

Eric BREHAULT – Plone Conference 2013
ZODB is good at storing objects

- Plone contents are objects,
- we store them in the ZODB,
- everything is fine, end of the story.
But what if …

… we want to store non-contentish records?

Like polls, statistics, mail-list subscribers, etc., or any **business-specific structured data**.
Store them as contents anyway

That is a powerfull solution.

But there are 2 major problems...
Problem 1: You need to manage a secondary system

- you need to **deploy** it,
- you need to **backup** it,
- you need to **secure** it,
- etc.
Problem 2: I hate SQL

No explanation here.
I think I just cannot digest it...
How to store many records in the ZODB?

- Is the ZODB strong enough?
- Is the ZCatalog strong enough?
My grandmother often told me

"If you want to become stronger, you have to eat your soup."
Where do we find a good soup for Plone?

In a super souper!!!
souper.plone and souper

- It provides both **storage** and **indexing**.
- Record can store any persistent pickable data.
- Created by **BlueDynamics**.
- Based on ZODB BTrees, node.ext.zodb, and repoze.catalog.
Add a record

```python
>>> soup = get_soup('mysoup', context)
>>> record = Record()
>>> record.attrs['user'] = 'user1'
>>> record.attrs['text'] = u'foo bar baz'
>>> record.attrs['keywords'] = [u'1', u'2', u'ü']
>>> record_id = soup.add(record)
```
Record in record

```python
>>> record['homeaddress'] = Record()
>>> record['homeaddress'].attrs['zip'] = '6020'
>>> record['homeaddress'].attrs['town'] = 'Innsbruck'
>>> record['homeaddress'].attrs['country'] = 'Austria'
```
Access record

```python
>>> from souper.soup import get_soup
>>> soup = get_soup('mysoup', context)
>>> record = soup.get(record_id)
```
Query

```python
>>> from repoze.catalog.query import Eq, Contains
>>> [r for r in soup.query(Eq('user', 'user1')
    & Contains('text', 'foo'))]
[<Record object 'None' at ...>]
```

or using CQE format

```python
>>> [r for r in soup.query("user == 'user1' and 'foo' in text")]
[<Record object 'None' at ...>]
```
souper

- a Soup-container can be moved to a specific ZODB mount-point,
- it can be shared across multiple independent Plone instances,
- souper works on Plone and Pyramid.
Plomino & souper

- we use Plomino to build **non-content oriented apps** easily,
- we use souper to store **huge amount** of application data.
Plomino data storage

Originally, documents (=record) were ATFolder.

Capacity about 30 000.
Plomino data storage

Since 1.14, documents are pure CMF.

Capacity **about 100 000**.

Usally the Plomino ZCatalog contains a lot of indexes.
Plomino & souper

With souper, documents are just soup records.

Capacity: **several millions**.
Typical use case

• Store 500 000 addresses,
• Be able to query them in full text and display the result on a map.

Demo
What is the limit?

Can we import Wikipedia in souper?

Demo with 400 000 records
Demo with 5,5 millions of records
Conclusion

- Usage performances are good,
- Plone performances are not impacted.

Use it!
Thoughts

- What about a REST API on top of it?
- Massive import is long and difficult, could it be improved?
Makina Corpus

For all questions related to this talk, please contact Éric Bréhault
eric.brehault@makina-corpus.com

Tel : +33 534 566 958

www.makina-corpus.com