НАБИС ФОНДАЦИЯ НАЦИОНАЛНА АКАДЕМИЧНА БИБЛИОТЕЧНО-ИНФОРМАЦИОННА СИСТЕМА

Проблеми и решения около агрегирането на фотографски колекции

Национален форум "Преображенията на Б-пространството", Варна, 21-22.03.2013

Евгени Димитров

НАБИС ФОНДАЦИЯ НАЦИОНАЛНА АКАДЕМИЧНА БИБЛИОТЕЧНО-ИНФОРМАЦИОННА СИСТЕМА

Фондация НАБИС беше основана в 2009 година от:

• Централната библиотека на Българската академия на науките,

• Софийския университет "Св. Климент Охридски" и

• Американския университет в България

с поддръжката на фондация Америка за България.

Основна цел на фондацията е да създаде своден каталог на българските научни и университетски библиотеки.

НАБИС ФОНДАЦИЯ НАЦИОНАЛНА АКАДЕМИЧНА БИБЛИОТЕЧНО-ИНФОРМАЦИОННА СИСТЕМА

Тук ще стане дума за два проблема, които възникнаха при участието ни в проекта EuropeanaPhotography.

Както вече е ставало дума, НАБИС няма собствена фотографска колекция и участието ни се основава на сътрудничеството на други организации и частни колекционери, които предоставят фотографии за проекта.

Те предоставят фотографии и описания – метаданни.

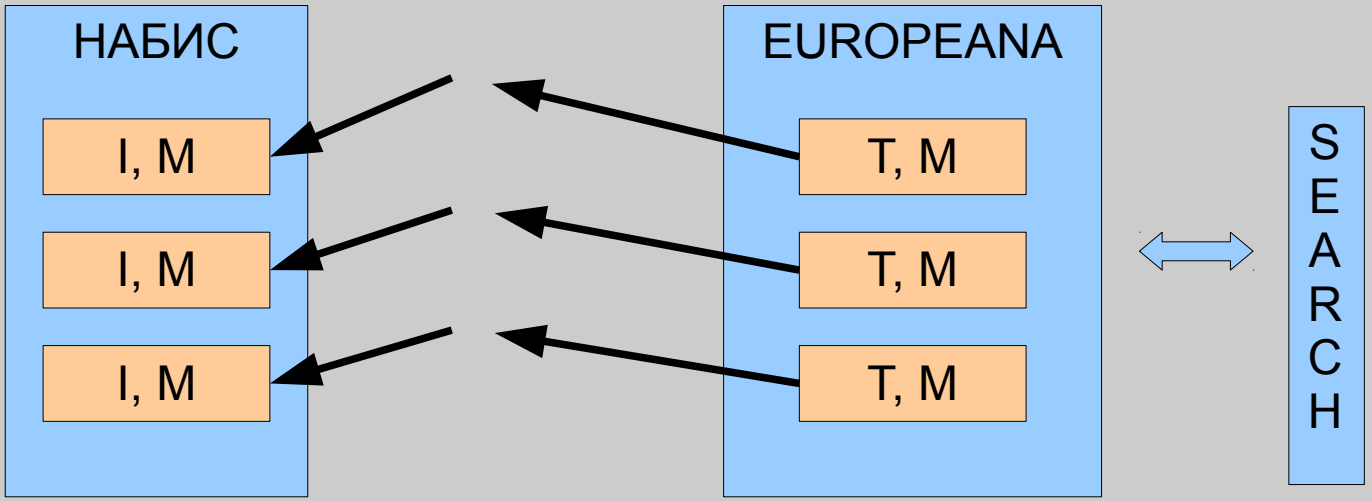
НАБИС помага първо в процеса на дигитализирането на фотографиите.

След това НАБИС поставя фотографиите на своите сървери и ще ги поддържа достъпни за Europeana за най-малко пет години след приключването на проекта.

Първият проблем, за който искам да разправя, е свързан с тази многогодишна поддръжка.

Проблем първи

Както е ставало дума и преди, когато фотографиите са публикувани в Europeana, върху сървърите на Europeana се запазват само линкове, които сочат към самите фотографии – които в нашия случай са върху сървърите на НАБИС.



Линковете са построени според това как нашата система показва фотографиите на външния свят и те изглеждат горе-долу така:

http://my_server/my_system?id=111277

Тук две неща са малко произволни и временни – името на системата и системния номер на обекта. Номерът на обекта е, като правило, уникален номер, който се дава на обекта при въвеждането му в системата. Нещо като пореден номер.

Ако след година или две решим да сменим системата си – дигиталния си архив – даже след най-успешно прехвърляне на данните от старата система в новата, най-вероятно системният номер на обекта ще бъде различен. Новият линк ще бъде нещо като:

http://my_server/my_new_system?id=102044

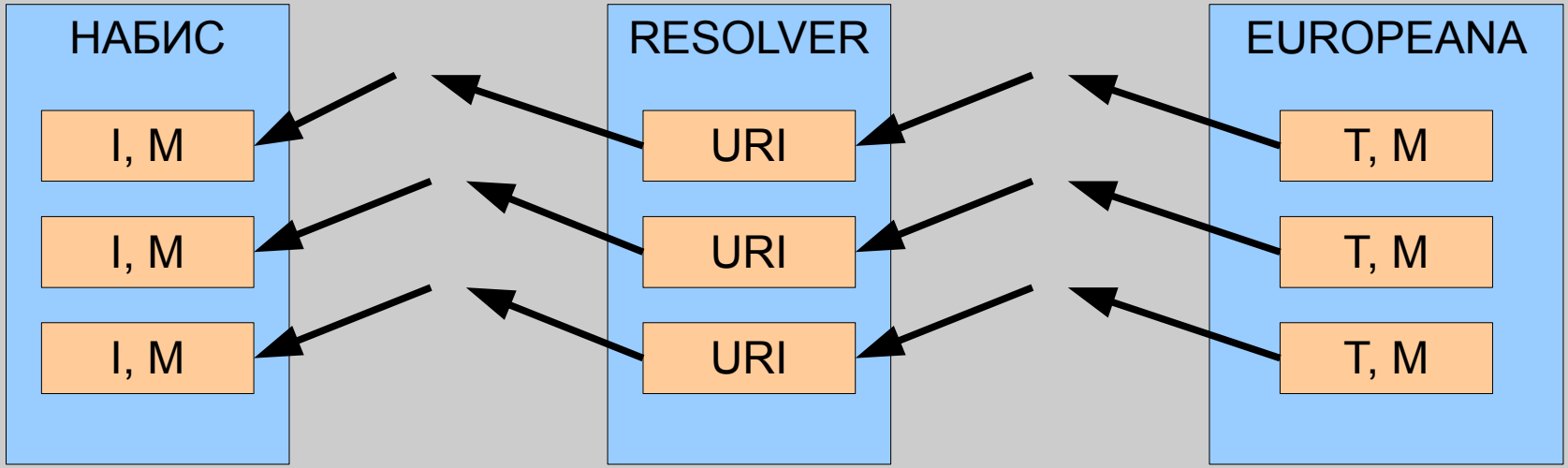
Което значи, че линкът, който е върху сървърите на Europeana, вече няма да работи.

Това е известен проблем и в такива случаи се препоръчва да се използват "persistent identifiers". На всеки обект се дава постоянно име – например използвайки сигнатурата. Тогава може да се направи таблица:

PID	URI
820_343	http://my_server/my_system?id=111277

В едната колонка е "постоянния идентификатор", в другата колонка – текущия линк.

Тази таблица се поставя в един трети сървер и се получава следната картина:



Europeana има линк към resolver-а – например:

`http://some_server/resolver/820_343`

А resolver-ът пренасочва към текущия линк:

`http://my_server/my_system?id=111277`

По този начин, ако ни се наложи да правим смяна на системата, ще трябва само да променим таблицата в resolver-а.

Остава за всеки конкретен случай да се реши къде да бъде той. Има организации, които предлагат тази услуга срещу заплащане. В някои страни има такива национални системи. А една организация – ако реши – може сама да поддържа за себе си такъв resolver. В края на краищата, това е просто една програма, която може да се изпълнява на някой от съществуващите сървери.

Това, всъщност, направихме и ние. И вече проверихме, че това решение работи. Защото, макар и да не е започнало публикуването във Europeana чрез проекта EuropeanaPhotography, ние вече публикувахме във Europeana известен брой обекти от сегашната ни система чрез проекта LinkedHeritage.

Това е всичко за първия проблем.

Проблем втори

Ние събираме фотографии от много организации и частни колекционери. Бихме искали да избегнем повторения – да не въвеждаме в Europeana еднакви фотографии.

Най-първата стъпка беше да групираме фотографиите тематично, така че ако някоя от новите фотографии ни се струва позната, да не преравяме всички стари, а само една част. Това общо взето работи.

След това опитахме и един по-модерен подход, за който сега искам да кажа.

Идеалното решение би било да подадем пакет от нови фотографии на една програма и тя да каже кои от новите съвпадат с кои от старите.

И в най-простия вариант, това са сериозни програми – не са нещо, което човек да напише от начало и до край. От друга страна не беше оправдано да купуваме скъп софтуер – това не беше най-големия ни проблем, а и бюджетът по проекта не е голям.

Направихме няколко опита, като използвахме една open source библиотека – LIRE – Lucene Image Retrieval – на Mathias Lux.

Mathias Lux преподава в университета на Клагенфурт, Австрия и има редица публикации в тази област.

Опит първи

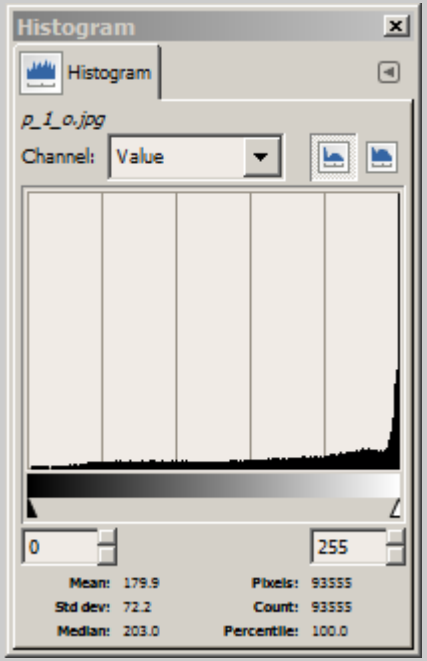
Първият опит беше с използване на "глобални свойства" на фотографията. Това е подход, когато се опитваме да опишем фотографията с неголям брой числа.

Това е най-простия подход и исторически най-ранен.

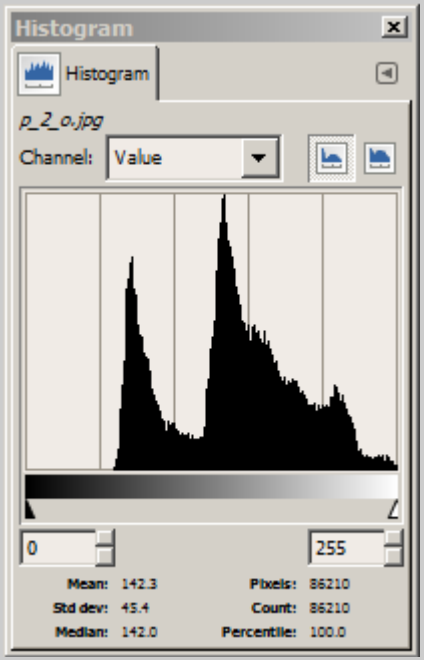
В най-проста форма такова числено описание на фотографията е хистограмата.

Хистограма

Прави се много просто – преброяват се колко точки са съвсем черни, колко точки са съвсем бели и колко точки са във всяка степен на сивото.



Различните фотографии имат общо взето различни хистограми.



Оказа се, обаче, че има твърде много фотографии с приличащи си хистограми и този подход не помага много за намиране на дублирани фотографии.

Донякъде проблемът идва и от това, че нашите фотографии са черно-бели – иначе можехме да броим не само в степени на сивото, но и в степени на червеното, степени на синьото и т.н.

Така или иначе, преминахме към втори опит. Идеята на алгоритъма във втория опит е да се намерят във фотографията особени точки, да се групират по определен начин и тези групи да се интерпретират като ключови думи (visual words).

Намират се тези "ключови думи" във старите фотографии и се индексират. Когато дойде нова фотография, тогава се намират "ключовите думи" в нея и се търсят стари фотографии със същите "ключови думи".

Опит втори

Започва се просто – фотографията се "размива".

После от истинската фотография се изважда "размитата" фотография и остават "особени точки" – контури.

Нататък от тези точки се правят ключовите думи (или "визуални думи").

Тук, обаче, математиката става по-тежка и не съм я проследявал до край. Така или иначе, това го има реализирано в библиотеката, която използвахме.

Крайният резултат от втория опит е, че този подход е практически почти използваем. В какъв смисъл?

Когато идват нови фотографии, нашата програма намира за всяка от тях трите стари, които най-много и приличат.

За всяка нова фотография се получава нещо такова:

```
CI3305.jpg
    0.22120517::TarnovoRegLib_Daniela_images_012
    0.06856811::PlevenDVIM_Daniela_B-V-0046
    0.0::NIM01_Daniela_8690
```

Идеята е, че колкото е по-голямо числото отпред, толкова е по-голяма приликата.

Резултатът беше, че за почти всички нови фотографии, които дублираха някоя стара, старата (дублираната) беше в тази тройка – обикновенно на първо място.

Но за много нови фотографии, които не дублираха стари, числата отпред бяха много големи – т.е. на тези числа не може да се разчита – те да покажат, че новата фотография дублира стара.

И така, заключението за втория поход за разпознаване на дублирани фотографии е, че:

- Той не може да бъде използван като напълно автоматичено средство.
- Той, обаче, може да показва малки изображени на новата фотография и на трите стари, които най-много и приличат, и да предлага на човек да взема окончателно решение.

Това е всичко.

Адрес за контакти: evgeni.dimitrov@nalis.bg

Има ли въпроси?

***