# Prime sperimentazioni d'indicizzazione [semi]automatica alla BNCF

Maria Grazia Pepe - Elisabetta Viti
(Biblioteca nazionale centrale di Firenze)

6. Incontro ISKO Italia
Firenze 20 maggio 2013

## SOMMARIO

- Partners ed obiettivi del progetto
- Indicizzazione automatica: definizione, utenti e metodologie
- Funzione delle parole/frasi chiave e procedure per l'estrazione automatica
- Indicizzazione umana (assegnata) vs. Indicizzazione automatica (derivata)
- Dalla teoria alla pratica: le prime sperimentazioni
- Creazione dei modelli di apprendimento
- Primi tentativi d'indicizzazione
- Analisi dei risultati e problemi aperti

## Partners e obiettivi del progetto

**Partners**:

- Biblioteca nazionale centrale di Firenze (BNCF)
- Casalini libri
- @Cult, azienda di progettazione e sviluppo informatico di Roma

**Obiettivi:**

1. Impiegare il thesaurus del Nuovo soggettario nell'indicizzazione automatica di risorse digitali con lo spirito di adeguare strumenti catalografici tradizionali al crescente sviluppo del mondo dell'informazione
2. Ridurre i costi della catalogazione anche razionalizzando risorse umane e finanziarie

Il lavoro è consistito di varie fasi, alcune prettamente informatiche ed altre maggiormente biblioteconomiche. Fra i principali step:

- Estrazione di parole/frasi chiave pertinenti e controllate da documenti digitali in full-text, depositati presso la BNCF (L. 106/2004, D.P.R. 252/2006)
- Assegnazione di parole/frasi chiave ai documenti digitali in full-text

## Indicizzazione automatica: definizione

Quando si parla d'indicizzazione automatica o semiautomatica ci si riferisce all'impiego di algoritmi e di alcune tecniche che possono sostituire o integrare l'intervento umano nell'analisi e indicizzazione di documenti espressi nel linguaggio naturale, in un formato leggibile dalla macchina, così che i termini d'indicizzazione estratti, sulla base della loro rilevanza semantica, siano associati al loro contenuto

## Chi usa metodi d'indicizzazione automatica

- Motori di ricerca in Internet
- Database di fulltext (es. database di periodici)
- Database bibliografici: per es. OPAC

## Metodi d'indicizzazione automatica

- assegnazione ai documenti in formato digitale dei soggetti presenti all'interno di record bibliografici relativi ai corrispettivi documenti in formato cartaceo
- estrazione e assegnazione di parole/frasi chiave (combinazione di parole usate come stringhe possibili di ricerca) con il supporto di un vocabolario controllato
- estrazione di parole/frasi chiave per la generazione automatica di indici e loro assegnazione al documento senza il supporto di un vocabolario controllato

## Funzione delle parole o frasi chiave

Le parole o frasi chiave possono essere uno strumento importante per:

- riassumere i contenuti concettuali dei documenti in modo conciso
- raggruppare i documenti in base al livello di sovrapposizione delle parole/frasi, impiegandole come parametro di somiglianza
- ricercare il contenuto concettuale di risorse digitali

Queste funzioni non sono poi così diverse da quelle assolte dalle intestazioni per soggetto nell'indicizzazione tradizionale!

## Procedura per l'estrazione automatica di parole chiave

- estrazione dei metadati assegnati esplicitamente da autori, editori, biblioteche oppure creati automaticamente
- estrazione di parole assegnate da autori/editori all'interno dei documenti
- separazione dei termini dai “non termini”
- identificazione di termini rilevanti (sia semplici che composti) all'interno del testo

## Metodo per l'individuazione ed esclusione dei "non termini"

- Impiego di filtri linguistici (POS tagging; liste di stopwords)
- Procedura di stemming
  - riduzione della parole alla radice
  - esclusione di parole/frasi che ricorrono solo una volta
- Individuazione ed esclusione di parole/frasi non significative per un determinato dominio disciplinare

## Strumenti per la valutazione del "peso semantico" dei termini

Misure statistiche che spaziano dal computo della normale frequenza a quello della frequenza relativa, come per esempio Term *Frequency/Inverse Document Frequency (TF/IDF)*, fino a metodi più sofisticati come il metodo *C-NC Value* o ad associazioni di misure come *Log likelihood*

| Indicizzazione umana (assegnata) | Indicizzazione automatica (derivata) |
|---|---|
| **Sistema più costoso** in termini di tempo risorse finanziarie e risorse umane | **Sistema più economico** in termini di tempo risorse finanziarie e risorse umane |
| **Fatta da Catalogatori** con formazione specifica e aggiornamento professionale continuo | **Fatta da sistemi complessi** (*software*) non completamente standardizzati |
| Uso di regole e standard | --------- |
| **Processo cognitivo che riflette la cultura individuale dell'indicizzatore**<br>• Analisi del testo<br>• Comprensione del testo e delle parole<br>• Interpretazione e individuazione del contenuto concettuale (tema di base) | **Processo meccanico su base statistica**<br>• Analisi del testo<br>• Nessuna comprensione e interpretazione del testo |
| **Decodifica tramite l'uso di un linguaggio documentario e di un vocabolario controllato, con l'obiettivo di condividere l'informazione con l'utente** | ---------- |
| **Linguaggio controllato** | **Linguaggio libero** |
| **I termini d'indicizzazione vengono attinti da un vocabolario controllato (soggettario, thesaurus, etc.)**<br>**La scelta dei termini d'indicizzazione è basata su conoscenza e comprensione** | **I termini d'indicizzazione sono estratti in automatico** grazie ad una ricerca libera sul *full text* con il supporto o meno di un vocabolario controllato<br>**La scelta dei termini di indicizzazione è basata solo su procedimenti statistici (es. occorrenze in porzioni di testo)** |
| **Livello alto di Precisione**<br>**Livello basso di richiamo** | **Livello basso di Precisione**<br>individua parole con un grado maggiore di occorrenza a discapito di parole/espressioni meno comuni ma comunque significative<br>**Livello alto di Richiamo**<br>individua anche i sinonimi (qualora non vengano impiegati strumenti di controllo terminologico) oppure indicizza parole usate dagli autori che possono risultare semanticamente più generiche |

# Fasi del progetto

**Fase 1: Dicembre 2010-Ottobre 2011**

- Individuazione della tipologia dei documenti e relativi metadati: tesi di dottorato acquisite in BNCF, dagli archivi aperti delle Università italiane (circolare del MIUR n.1746 del 20 luglio 2007)
- scelta del thesaurus del Nuovo soggettario come componente base per le procedure d'estrazione
- definizione di procedure e flussi di lavoro

## Fasi del progetto

**Fase 2: Novembre 2011-Dicembre 2012**

- Realizzazione del software per l'estrazione e l'associazione di parole chiave:
  - software di base Open Source
  - realizzazione del software *Keyword Indexer* (KI)
  - analisi del Nuovo soggettario (NS) in versione SKOS/RDF
- sperimentazione:
  - modello di apprendimento multidisciplinare
  - modello di apprendimento in uno specifico dominio disciplinare

## Modello di apprendimento

- set di documenti digitali significativi di dominio
- set di metadati associati ai documenti in full text
- parametri aggiuntivi :
  - vocabolario controllato
  - stemming
  - lingua.

## Modello di apprendimento

Ogni modello di apprendimento serve per elaborare una sorta di distribuzione statistica dei termini presenti nel set di documenti campione a cui è attribuito un “peso” in funzione di alcuni parametri quali:

- presenza all'interno dei metadati
- percentuale dello spazio che nel documento precede la prima occorrenza del termine
- lunghezza della frase
- TF / IDF
- presenza o meno del termine all'interno del vocabolario controllato;

## Sperimentazione: creazione dei modelli di apprendimento

Modello A:

- 200 tesi in formato pdf di ambito multidisciplinare ;
- parole chiave estratte dai metadati semantici associati alle tesi;
- thesaurus del Nuovo soggettario in formato SKOS/RDF.

Modello B:

- 100 tesi in formato pdf di uno specifico dominio disciplinare (MIUR area 8 - Ingegneria civile e Architettura);
- parole chiave estratte dai metadati semantici associati alle tesi;
- thesaurus del Nuovo soggettario in formato SKOS/RDF.

## Procedure seguite

- *analisi dei metadati semantici:* creazione di una tabella di corrispondenza tra la decodifica verbale dei codici di classificazione disciplinare MIUR ed i termini del NS;
- *selezione della lingua:* esclusione dei documenti in cui la lingua dell'abstract differisce da quella del testo (altrimenti necessario il supporto di un vocabolario controllato multilingue);
- *formato del testo:* esclusione dei documenti con un contenuto elevato di elementi grafici, formule matematiche ecc...

## Modelli di apprendimento con parole chiave attribuite anche automaticamente

A partire dai due modelli di apprendimento appena descritti ne sono stati creati due ulteriori (modelli A1 e B1) ampliando l'elenco delle parole chiave estratte dai metadati semantici con i termini del NS che sono “non preferiti” ma hanno una relazione di equivalenza (cioè di sinonimia) con termini “preferiti”

# Documenti sperimentalmente indicizzati in modalità automatica

1. Losasso M.,D'Ambrosio V., *Eco-quartieri e Social Housing nelle esperienze nord europee*, "Techne" 4(2012)

2. Creazza A., Dallari F., Leone F., *Analisi delle esigenza logistiche e sviluppo di soluzioni operative per Expo 2015*, "LIUC Papers", serie Tecnologia (ott. 2012)

# Risultati (1): modello di apprendimento A

*Eco-quartieri e Social Housing nelle esperienze nord europee*


```
−<response status-code="1">
 −<document filename="11489-20005-1-SM.pdf">
  −<keyword-indexing-request id="113673188366 79" date="02/05/2013 12:36:08">
     <description>tesi_2013</description>
     <training-model-name>tesi_2013_A</training-model-name>
     <vocabulary-name>NS</vocabulary-name>
     <language-code>it</language-code>
     <use-stemming>false</use-stemming>
    −<keywords>
       <keyword uri="http://purl.org/bncf/tid/8182" tf="0.00437158" idf="1.49664" tfXidf="0.0065426817" spread="0.92552">Urbanistica</keyword>
       <keyword uri="http://purl.org/bncf/tid/17555" tf="0.00218579" idf="4.61016" tfXidf="0.010076841" spread="0.0214259">Paesi
       scandinavi</keyword>
       <keyword uri="http://purl.org/bncf/tid/17372" tf="0.00218579" idf="3.00072" tfXidf="0.0065589435" spread="0.222931">Centri
       commerciali</keyword>
       <keyword uri="http://purl.org/bncf/tid/48233" tf="0.00327869" idf="2.73836" tfXidf="0.008978234" spread="0.300089">Valutazione di impatto
       ambientale</keyword>
       <keyword uri="http://purl.org/bncf/tid/8083" tf="0.00218579" idf="2.35887" tfXidf="0.0051559946" spread="0.08328">Spazi
       pubblici</keyword>
       <keyword uri="http://purl.org/bncf/tid/1527" tf="0.00218579" idf="1.47466" tfXidf="0.0032232972" spread="0.20023">Critica</keyword>
       <keyword uri="http://purl.org/bncf/tid/5177" tf="0.00218579" idf="1.41148" tfXidf="0.0030851988" spread="0.251626">Identità</keyword>
       <keyword uri="http://purl.org/bncf/tid/12463" tf="0.00218579" idf="1.39128" tfXidf="0.003041046" spread="0.183523">Funzionali</keyword>
       <keyword uri="http://purl.org/bncf/tid/15074" tf="0.00327869" idf="1.37148" tfXidf="0.0044966578" spread="0.3063388">Temi</keyword>
       <keyword uri="http://purl.org/bncf/tid/14426" tf="0.00327869" idf="1.37148" tfXidf="0.0044966578" spread="0.139778">Costi</keyword>
     </keywords>
   </keyword-indexing-request>
  </document>
</response>
```

# Risultati (1): modello di apprendimento A1

*Eco-quartieri e Social Housing nelle esperienze nord europee*


```
−<response status-code="1">
  −<document filename="11489-20005-1-SM.pdf">
    −<keyword-indexing-request id="11367318836635" date="02/05/2013 11:32:29">
       <description>tesi_2013</description>
       <training-model-name>tesi_2013_I</training-model-name>
       <vocabulary-name>NS</vocabulary-name>
       <language-code>it</language-code>
       <use-stemming>false</use-stemming>
      −<keywords>
         <keyword uri="http://purl.org/bncf/tid/8182" tf="0.00437158" idf="1.49664" tfXidf="0.0065426817" spread="0.92552">Urbanistica</keyword>
         <keyword uri="http://purl.org/bncf/tid/1218" tf="0.0240437" idf="0.837397" tfXidf="0.020134123" spread="0.8706801">Città</keyword>
         <keyword uri="http://purl.org/bncf/tid/1133" tf="0.00218579" idf="1.47466" tfXidf="0.0032232972" spread="0.248438">Aziende</keyword>
         <keyword uri="http://purl.org/bncf/tid/11274" tf="0.00327869" idf="1.26025" tfXidf="0.004131969"
         spread="0.512434">Organizzazione</keyword>
         <keyword uri="http://purl.org/bncf/tid/5063" tf="0.00327869" idf="1.09861" tfXidf="0.0036020016" spread="0.753348">Gestione</keyword>
         <keyword uri="http://purl.org/bncf/tid/7789" tf="0.00655738" idf="0.908856" tfXidf="0.0059597143" spread="0.585894">Realtà</keyword>
         <keyword uri="http://purl.org/bncf/tid/17555" tf="0.00218579" idf="4.61016" tfXidf="0.010076841" spread="0.0214259">Paesi
         scandinavi</keyword>
         <keyword uri="http://purl.org/bncf/tid/37101" tf="0.00218579" idf="3.35739" tfXidf="0.0073385495" spread="0.225864">Centri di
         ricerca</keyword>
         <keyword uri="http://purl.org/bncf/tid/44412" tf="0.00327869" idf="3.22386" tfXidf="0.010570037" spread="0.342431">Biogas</keyword>
         <keyword uri="http://purl.org/bncf/tid/22093" tf="0.00218579" idf="3.00072" tfXidf="0.0065589435"
         spread="0.002551">Prestazione</keyword>
       </keywords>
     </keyword-indexing-request>
   </document>
 </response>
```


# Risultati (1) : modello di apprendimento B

*Eco-quartieri e Social Housing nelle esperienze nord europee*


```
−<response status-code="1">
  −<document filename="11489-20005-1-SM.pdf">
    −<keyword-indexing-request id="11366358640348" date="29/04/2013 15:43:16">
       <description>tesi_2013</description>
       <training-model-name>tesi_2013_B</training-model-name>
       <vocabulary-name>NS</vocabulary-name>
       <language-code>it</language-code>
       <use-stemming>false</use-stemming>
      −<keywords>
         <keyword uri="http://purl.org/bncf/tid/17555" tf="0.00272851" idf="4.14313" tfXidf="0.011304571" spread="0.0214259">Paesi
         scandinavi</keyword>
         <keyword uri="http://purl.org/bncf/tid/37101" tf="0.00272851" idf="3.04452" tfXidf="0.008307003" spread="0.225864">Centri di
         ricerca</keyword>
         <keyword uri="http://purl.org/bncf/tid/48233" tf="0.00409277" idf="2.5337" tfXidf="0.010369851" spread="0.300089">Valutazione di impatto
         ambientale</keyword>
         <keyword uri="http://purl.org/bncf/tid/8083" tf="0.00272851" idf="1.94591" tfXidf="0.005309435" spread="0.08328">Spazi
         pubblici</keyword>
         <keyword uri="http://purl.org/bncf/tid/630" tf="0.00272851" idf="1.1474" tfXidf="0.0031306923" spread="0.012116">Costruzioni</keyword>
         <keyword uri="http://purl.org/bncf/tid/8182" tf="0.00545703" idf="1.1474" tfXidf="0.006261396" spread="0.92552">Urbanistica</keyword>
         <keyword uri="http://purl.org/bncf/tid/24709" tf="0.0095498" idf="1.1474" tfXidf="0.01095744" spread="0.7413596">Obiettivi</keyword>
         <keyword uri="http://purl.org/bncf/tid/8468" tf="0.00409277" idf="1.12271" tfXidf="0.004594994" spread="0.237597">Principi</keyword>
         <keyword uri="http://purl.org/bncf/tid/36785" tf="0.00272851" idf="1.09861" tfXidf="0.0029975683" spread="0.645963">Stato
         nazionale</keyword>
         <keyword uri="http://purl.org/bncf/tid/28443" tf="0.00682128" idf="1.09861" tfXidf="0.007493926"
         spread="0.7204442">Accrescimento</keyword>
       </keywords>
     </keyword-indexing-request>
   </document>
 </response>
```

# Risultati (1): modello di apprendimento B1

*Eco-quartieri e Social Housing nelle esperienze nord europee*


```
<response status-code="1">
  <document filename="11489-20005-1-SM.pdf">
    <keyword-indexing-request id="11367318836643" date="02/05/2013 11:36:54">
      <description>tesi_2013</description>
      <training-model-name>tesi_2013_L</training-model-name>
      <vocabulary-name>NS</vocabulary-name>
      <language-code>it</language-code>
      <use-stemming>false</use-stemming>
      <keywords>
        <keyword uri="http://purl.org/bncf/tid/1218" tf="0.0240437" idf="0.616774" tfXidf="0.014829529" spread="0.8706801">Città</keyword>
        <keyword uri="http://purl.org/bncf/tid/8182" tf="0.00437158" idf="1.1474" tfXidf="0.005015951" spread="0.92552">Urbanistica</keyword>
        <keyword uri="http://purl.org/bncf/tid/17555" tf="0.00218579" idf="4.14313" tfXidf="0.009056012" spread="0.0214259">Paesi
        scandinavi</keyword>
        <keyword uri="http://purl.org/bncf/tid/37101" tf="0.00218579" idf="3.22684" tfXidf="0.0070531946" spread="0.225864">Centri di
        ricerca</keyword>
        <keyword uri="http://purl.org/bncf/tid/22093" tf="0.00218579" idf="3.22684" tfXidf="0.0070531946"
        spread="0.002551">Prestazione</keyword>
        <keyword uri="http://purl.org/bncf/tid/32216" tf="0.00437158" idf="3.22684" tfXidf="0.014106389"
        spread="0.7621478">Rendimento</keyword>
        <keyword uri="http://purl.org/bncf/tid/467" tf="0.00327869" idf="3.04452" tfXidf="0.009982037" spread="0.086723">Sport</keyword>
        <keyword uri="http://purl.org/bncf/tid/44412" tf="0.00327869" idf="3.04452" tfXidf="0.009982037" spread="0.342431">Biogas</keyword>
        <keyword uri="http://purl.org/bncf/tid/29628" tf="0.00218579" idf="2.89037" tfXidf="0.0063177417" spread="0.343324">Recidive</keyword>
        <keyword uri="http://purl.org/bncf/tid/6742" tf="0.00218579" idf="2.63906" tfXidf="0.005768431" spread="0.166433">Marketing</keyword>
      </keywords>
    </keyword-indexing-request>
  </document>
</response>
```


# Risultati (2): modello di apprendimento A

*Analisi delle esigenza logistiche e sviluppo di soluzioni operative per Expo 2015*


```
<response status-code="1">
  <document filename="253.pdf">
    <keyword-indexing-request id="11367318836739" date="03/05/2013 17:16:43">
      <description>tesi_2013</description>
      <training-model-name>tesi_2013_A</training-model-name>
      <vocabulary-name>NS</vocabulary-name>
      <language-code>it</language-code>
      <use-stemming>false</use-stemming>
      <keywords>
        <keyword uri="http://purl.org/bncf/tid/2807" tf="0.00356718" idf="5.3033" tfXidf="0.018917825" spread="0.0645563">Esposizioni
        universali</keyword>
        <keyword uri="http://purl.org/bncf/tid/7184" tf="0.00118906" idf="4.61016" tfXidf="0.0054817568" spread="0.121907">Provvedimenti
        cautelari</keyword>
        <keyword uri="http://purl.org/bncf/tid/33258" tf="0.00535077" idf="4.20469" tfXidf="0.022498328" spread="0.9568676">Grandi
        eventi</keyword>
        <keyword uri="http://purl.org/bncf/tid/14499" tf="0.00178359" idf="3.35739" tfXidf="0.005988207" spread="0.9358234">Olimpiadi</keyword>
        <keyword uri="http://purl.org/bncf/tid/2112" tf="0.00118906" idf="2.53072" tfXidf="0.003009178" spread="0.1562237">Enti
        pubblici</keyword>
        <keyword uri="http://purl.org/bncf/tid/7111" tf="0.00178359" idf="1.49664" tfXidf="0.0026693922" spread="0.353114">Logica</keyword>
        <keyword uri="http://purl.org/bncf/tid/26866" tf="0.00237812" idf="1.49664" tfXidf="0.0035591896"
        spread="0.4251585">Localizzazione</keyword>
        <keyword uri="http://purl.org/bncf/tid/650" tf="0.00297265" idf="1.49664" tfXidf="0.004448987" spread="0.9058674">Giochi</keyword>
        <keyword uri="http://purl.org/bncf/tid/13097" tf="0.00118906" idf="1.47466" tfXidf="0.0017534592" spread="0.002749">Membri</keyword>
        <keyword uri="http://purl.org/bncf/tid/1817" tf="0.00118906" idf="1.47466" tfXidf="0.0017534592"
        spread="0.602047">Classificazione</keyword>
      </keywords>
    </keyword-indexing-request>
  </document>
</response>
```

# Risultati (2): modello di apprendimento A1

*Analisi delle esigenza logistiche e sviluppo di soluzioni operative per Expo 2015*


```
<response status-code="1">
  <document filename="253.pdf">
    <keyword-indexing-request id="11367318836747" date="03/05/2013 17:17:14">
      <description>tesi_2013</description>
      <training-model-name>tesi_2013_I</training-model-name>
      <vocabulary-name>NS</vocabulary-name>
      <language-code>it</language-code>
      <use-stemming>false</use-stemming>
      <keywords>
        <keyword uri="http://purl.org/bncf/tid/7624" tf="0.00118906" idf="4.61016" tfXidf="0.0054817568" spread="0.331595">Farmacia</keyword>
        <keyword uri="http://purl.org/bncf/tid/12805" tf="0.00356718" idf="0.698135" tfXidf="0.002490373"
        spread="0.9224569">Territorio</keyword>
        <keyword uri="http://purl.org/bncf/tid/1133" tf="0.00297265" idf="1.47466" tfXidf="0.0043836483" spread="0.82093">Aziende</keyword>
        <keyword uri="http://purl.org/bncf/tid/11274" tf="0.00297265" idf="1.26025" tfXidf="0.003746282"
        spread="0.2051375">Organizzazione</keyword>
        <keyword uri="http://purl.org/bncf/tid/5063" tf="0.0178359" idf="1.09861" tfXidf="0.019594697" spread="0.985781">Gestione</keyword>
        <keyword uri="http://purl.org/bncf/tid/3143" tf="0.00118906" idf="5.3033" tfXidf="0.006305942" spread="0.611527">Elevatori</keyword>
        <keyword uri="http://purl.org/bncf/tid/30449" tf="0.00118906" idf="5.3033" tfXidf="0.006305942" spread="0.186652">Autisti</keyword>
        <keyword uri="http://purl.org/bncf/tid/49867" tf="0.00118906" idf="5.3033" tfXidf="0.006305942"
        spread="0.022182">Accreditamento</keyword>
        <keyword uri="http://purl.org/bncf/tid/37296" tf="0.00118906" idf="5.3033" tfXidf="0.006305942" spread="0.117641">Misure di
        sicurezza</keyword>
        <keyword uri="http://purl.org/bncf/tid/3807" tf="0.00118906" idf="5.3033" tfXidf="0.006305942" spread="0.170253">Imballaggi</keyword>
      </keywords>
    </keyword-indexing-request>
  </document>
</response>
```


# Risultati (2): modello di apprendimento B

*Analisi delle esigenza logistiche e sviluppo di soluzioni operative per Expo 2015*


```
<response status-code="1">
  <document filename="253.pdf">
    <keyword-indexing-request id="11367318836743" date="03/05/2013 17:17:02">
      <description>tesi_2013_icar</description>
      <training-model-name>tesi_2013_B</training-model-name>
      <vocabulary-name>NS</vocabulary-name>
      <language-code>it</language-code>
      <use-stemming>false</use-stemming>
      <keywords>
        <keyword uri="http://purl.org/bncf/tid/2112" tf="0.00118906" idf="2.5337" tfXidf="0.0030127214" spread="0.1562237">Enti
        pubblici</keyword>
        <keyword uri="http://purl.org/bncf/tid/24709" tf="0.00118906" idf="1.1474" tfXidf="0.0013643275" spread="0.038771">Obiettivi</keyword>
        <keyword uri="http://purl.org/bncf/tid/43140" tf="0.00118906" idf="1.1474" tfXidf="0.0013643275" spread="0.066926">Gittata
        cardiaca</keyword>
        <keyword uri="http://purl.org/bncf/tid/38081" tf="0.00178359" idf="1.1474" tfXidf="0.0020464913" spread="0.758365">Prodotto</keyword>
        <keyword uri="http://purl.org/bncf/tid/14009" tf="0.00297265" idf="1.1474" tfXidf="0.0034108185"
        spread="0.369608">Distribuzione</keyword>
        <keyword uri="http://purl.org/bncf/tid/7595" tf="0.00356718" idf="1.1474" tfXidf="0.0040929825"
        spread="0.72623">Pianificazione</keyword>
        <keyword uri="http://purl.org/bncf/tid/5648" tf="0.00178359" idf="1.12271" tfXidf="0.0020024544"
        spread="0.277467">Informazioni</keyword>
        <keyword uri="http://purl.org/bncf/tid/21329" tf="0.00653983" idf="1.12271" tfXidf="0.0073423325"
        spread="0.744146">Superficie</keyword>
        <keyword uri="http://purl.org/bncf/tid/28443" tf="0.00118906" idf="1.09861" tfXidf="0.0013063132"
        spread="0.0121338">Accrescimento</keyword>
        <keyword uri="http://purl.org/bncf/tid/982" tf="0.00178359" idf="1.07508" tfXidf="0.0019175019" spread="0.049293">Case</keyword>
      </keywords>
    </keyword-indexing-request>
  </document>
</response>
```

# Risultati (2): modello di apprendimento B1

*Analisi delle esigenza logistiche e sviluppo di soluzioni operative per Expo 2015*


```
−<response status-code="1">
 −<document filename="253.pdf">
  −<keyword-indexing-request id="11367318836751" date="03/05/2013 17:17:35">
     <description>tesi_2013_icar</description>
     <training-model-name>tesi_2013_L</training-model-name>
     <vocabulary-name>NS</vocabulary-name>
     <language-code>it</language-code>
     <use-stemming>false</use-stemming>
    −<keywords>
      <keyword uri="http://purl.org/bncf/tid/12805" tf="0.00356718" idf="0.677399" tfXidf="0.0024164042"
      spread="0.9224569">Territorio</keyword>
      <keyword uri="http://purl.org/bncf/tid/3143" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273" spread="0.611527">Elevatori</keyword>
      <keyword uri="http://purl.org/bncf/tid/30449" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273" spread="0.186652">Autisti</keyword>
      <keyword uri="http://purl.org/bncf/tid/7184" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273" spread="0.121907">Provvedimenti
      cautelari</keyword>
      <keyword uri="http://purl.org/bncf/tid/49867" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273"
      spread="0.022182">Accreditamento</keyword>
      <keyword uri="http://purl.org/bncf/tid/6629" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273" spread="5.68E-
      4">Importazione</keyword>
      <keyword uri="http://purl.org/bncf/tid/37296" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273" spread="0.117641">Misure di
      sicurezza</keyword>
      <keyword uri="http://purl.org/bncf/tid/3807" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273" spread="0.170253">Imballaggi</keyword>
      <keyword uri="http://purl.org/bncf/tid/7624" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273" spread="0.331595">Farmacia</keyword>
      <keyword uri="http://purl.org/bncf/tid/44451" tf="0.00118906" idf="4.83628" tfXidf="0.0057506273"
      spread="0.331785">Bancomat</keyword>
     </keywords>
    </keyword-indexing-request>
  </document>
</response>
```


# Analisi dei risultati

Nella fase di creazione del modello di apprendimento l'arricchimento in modo puramente automatico dei metadati può produrre rumore anziché raffinare il risultato (modello di apprendimento A1). Il contributo intellettuale dei bibliotecari per l'attribuzione delle parole chiave è quindi indispensabile e l'intero processo è definibile come **Indicizzazione semi-automatica.**

Dall'analisi della documentazione ci si è resi conto della necessità di usare non soltanto il Thesaurus del Nuovo soggettario ma anche le liste di autorità della BNI che contengono nomi propri e geografici.

## Problemi aperti

- È preferibile raffinare un modello di apprendimento multidisciplinare o creare tanti modelli specialistico settoriali per quanti sono i domini disciplinari di competenza della biblioteca?
- Nel primo caso quanto ampio deve essere il set di documenti analizzati?
- Nella fase di creazione del modello di apprendimento, è possibile prescindere dall'attribuzione delle parole chiave ricavate da metadati?
- Come risolvere i problemi di selezione della lingua utilizzando un vocabolario monolingue?
- Come risolvere i problemi di formato del testo (documenti con un'alta percentuale di grafici o formule)?

Grazie per la vostra Attenzione!!!

*Maria Grazia Pepe - Elisabetta Viti*
(Biblioteca Nazionale Centrale di Firenze)