Integrazione delle informazioni in rete: prospettive per le scienze della vita

Paolo Romano
Istituto Nazionale per la Ricerca sul Cancro
(paolo.romano@istge.it)

Sommario

- Gestione e integrazione delle informazioni
- Caratteristiche di dati e integrazione in biologia
- Metodi di integrazione
- Strumenti per l'integrazione
- Esempi

Gestione delle informazioni

- Archiviazione locale, isolata, mainframe
- Archiviazione locale, condivisa, LAN
- Archiviazione decentrata, non condivisa
- Archiviazione decentrata, condivisa → integrazione
- La rete come archivio locale, anche se decentrato → data GRID
- Quale integrazione: dati, applicazioni, processi

Integrazione: grado

- Sistemi strettamente integrati:
  - Dati: warehouse
  - Applicazioni: centralizzati, CORBA
  - Processi: servizi statici, ripetitivi
  - Integrazione: precoce o predefinita
  - Trasparenza: elevata
- Sistemi a integrazione dinamica:
  - Dati: decentrati, integrazione dinamica
  - Applicazioni: Web Services, Data GRID
  - Processi: servizi dinamici, adattati su richiesta
  - Integrazione: a richiesta o scoperta
  - Trasparenza: interattiva

Integrazione: longevità

- L'integrazione necessita di stabilità:
  - Standardizzazione
  - Obiettivi precisi
  - Buona definizione dei dati
  - Buona comprensione del dominio
- L'integrazione teme:
  - Specializzazione dei dati
  - Eterogeneità dei dati e dei sistemi
  - Rapida evoluzione dei dati
  - Spontaneità, sperimentalismo
  - Mancanza di obiettivi predefiniti

L’informazione biologica

- La ricerca biomedica produce una quantità sempre crescente di dati
- Alcuni settori, quali la genomica e la proteomica, contribuiscono alla realizzazione di banche dati di rilevanti dimensioni
- Altri settori emergenti, legati all’analisi delle mutazioni, ai polimorfismi, al metabolismo, e derivati da nuovi strumenti quali i microarray, contribuiranno anch’essi con quantità di dati ancora superiori

L'informazione biologica

- EMBL Data Library 73 (Dec 2002):
  - Sequenze: 20.857.746, Basi: 27.903.283.528
  - Dimensione: ~91 Gbyte, 20% in più 72
- GenBank Release 125.0 (Aug 2001):
  - Sequenze: 12.813.516, Basi: 13.543.364.296
  - Dimensione: 49,72 GByte

➢ Questa enorme quantità di dati può essere analizzata solo tramite software specialistici

Banche dati eterogenee

- Alcune banche dati sono gestite da pochi Centri (EBI, NCBI, DDBJ) in modo disomogeneo
- Molte banche dati sono sviluppate da singoli ricercatori o piccolo gruppi
- Banche dati secondarie, ottima qualità (annotazione estesa, controllo accurato)
- Banche dati specializzate: gene/genoma, organismo, malattia, bibliografiche

Banche dati distribuite

- Banche dati non centralizzate significano:
  - Database differenti
  - Strutture dati differenti
  - Informazioni differenti
  - Significati differenti
  - Metodi di distribuzione dati differenti

Software per analisi biologiche

- I software sviluppati sono spesso alla base dell'analisi
  - Analisi di sequenza
  - Predizione di strutture proteiche secondarie e terziarie
  - Predizione di introni/esoni
  - Analisi evoluzione molecolare
  - Ricostruzione sequenze

Software per analisi biologiche

- È necessario integrare questi software con le banche dati
  - Banche dati come input dei software
  - Risultati elaborazione/analisi come nuovi dati da memorizzare e analizzare

Integrazione delle banche dati

- L'integrazione delle banche dati è necessaria per
  - Ottenere una visione complessiva delle informazioni disponibili
  - Eseguire in un numero limitato di passaggi interrogazioni e/o analisi che coinvolgono più banche dati e software
  - Effettuare un reale data mining

Integrazione delle banche dati

- L'integrazione delle banche dati comporta
  - L'analisi e la definizione accurata e univoca degli "oggetti biologici" coinvolti
  - L'analisi dei dati disponibili
  - L'identificazione dei collegamenti tra informazioni presenti in banche dati diverse
  - La definizione e l'implementazione di formati comuni per l'interscambio delle informazioni

Specificità dell'integrazione

- In ambito biologico:
  - Le esigenze e gli obiettivi di ricerca evolvono rapidamente, seguendo le nuove acquisizioni
  - Una pre-analisi e riaggregazione delle informazioni è impossibile, perché dati e conoscenze teoriche cambiano rapidamente
  - La complessità delle informazioni rende molto difficile creare modelli validi in diversi ambiti

I metodi dell'integrazione

- Sintattici
  - Riferimenti reciproci (xrefs)
  - Descrizioni condivise (vocabolari)
- Semantici
  - Modelli a oggetti
  - Schemi relazionali
  - Ontologie

Riferimenti reciproci

- Da un record, a un record collegato di un'altra banca dati:
  - Link diretto, univoco, non reciproco
  - ID database remoto
  - Formati standardizzati
    - Life Science ID
    - Standard I3C
- Limitazioni:
  - Annotazione manuale
  - Significato del collegamento
  - Predefiniti

Descrizioni condivise

- Da un record ai record di un'altra banca dati tramite ricerca testuale:
  - Link implicito, reciproco
  - Determinabile automaticamente
  - Termine di vocabolario
  - Vocabolari standardizzati
- Limitazioni:
  - Diffusione di vocabolari condivisi nell'annotazione
  - Significato del collegamento
  - Necessità di definire l'ambito
  - Text mining

XML (eXtensible Markup Language)

- Linguaggio Markup per inserire semantica del documento:
  - Supera limiti HTML
  - Semplice definizione e implementazione tramite Document Type Definitions (DTDs)
  - Modulare, nuovi DTD utilizzano precedenti
- Utilizzabile da applicazioni software:
  - Well formed
  - Vocabolari standardizzati
  - Analisi sintattica ed estrazione dati automatiche

```
ID   AA415057   standard; RNA; EST; 337 BP.
AC   AA415057;
SV   AA415057.1
DT   27-OCT-1997 (Rel. 53, Created)
DT   14-DEC-1999 (Rel. 62, Last updated, Version 2)
DE   Mg0001 RCW Lambda Zap Express Library Pyricularia grisea cDNA clone RCW1
DE   5', mRNA sequence.
KW   EST.
OS   Magnaporthe grisea
OC   Eukaryota; Fungi; Ascomycota; Pezizomycotina; Sordariomycetes;
OC   Sordariomycetes incertae sedis; Magnaporthaceae; Magnaporthe.
RN   [1]
RP   1-337
RA   Wu S.-C., Bernstein B.D., Darvill A.G., Albersheim P.;
RT   "Expressed sequence tags of the rice blast fungus grown on rice cell
RT   walls";
RL   Unpublished.
DR   UNILIB; 863; 863.
CC   Contact: Sheng-Cheng Wu
CC   CCRC
CC   University of Georgia
CC   220 Riverbend Road, Athens, GA 30602-4712, USA
CC   Tel: 706 542 4446
CC   Fax: 706 542 4412
CC   Email: wusc@bscr.uga.edu
CC   Seq primer: T3.
```

```
FH   Key             Location/Qualifiers
FH
FT   source          1..337
FT                   /db_xref="taxon:148305"
FT                   /db_xref="UNILIB:863"
FT                   /note="Vector: Lambda Zap; Messenger RNAs prepared from
FT                   Magnaporthe grisea grown at  23C in the dark with constant
FT                   gyratory shaking (100 rpm)  in Vogel's medium containing
FT                   0.5% isolated rice cell walls  as the sole carbon source"
FT                   /organism="Magnaporthe grisea"
FT                   /strain="CP987"
FT                   /clone="RCW1"
FT                   /clone_lib="RCW Lambda Zap Express Library"
FT                   /tissue_type="Mycelium"
FT                   /dev_stage="Day 5 post-inoculation"
XX
SQ   Sequence 337 BP; 56 A; 111 C; 74 G; 96 T; 0 other;
     cttttttcaat cagcccgaga actcctggtt gggttttctg cctgttctga cagctacttg        60
     tcatcgcata gcccgttctt tggttccaga taccacaagc ctgggacatt gatttcccag       120
     caactctttc aaaatggtat tattagcctc ctcacgatcc ctcgcgcgtt cgcttggtcc       180
     ccttgcttca cgctctcgag cttttcagag cagtgctgct tcccgttccc tctcgactgc       240
     caccgctcgc ggccagggca aatccacaac tctcctgagg cccgcggccg ccacgaggac       300
     aagcaggttg ttgtcgactg gttcgcccct tcgtatt                                337
//
```

```
<interpro id="IPR000002">
  <name>FIZZY/CDC20 domain</name>
  <type>Domain</type>
  <abstract> This domain is found in proteins …</abstract>
  <examplelist>
    <example>
      <protein sptr_ac="Q12834" />Mammalian protein, p55CDC
    </example>
    <example>
      <protein sptr_ac="Q09649" />
    </example>
  </examplelist>
  <publist>
    <publication pub_id="PUB00006167">
     <authorlist>Shirayama M., Toth A., Galova M., Nasmyth K.</authorlist>
     <title>APC(Cdc20) promotes exit from mitosis by ……</title>
     <dbxref db="MEDLINE" dbkey="20110935" />
     <journal>Nature</journal>
     <location firstpage="203" lastpage="207" volume="402" />
     <year>1999</year>
    </publication>
  </publist>
  <memberlist>
    <dbxref db="PREFILE" dbkey="PS50218" name="FIZZY_DOMAIN" />
    <dbxref db="PRODOM" dbkey="PD004563" name="PD004563" />
  </memberlist>
</interpro>
```

```
<!ELEMENT interpro (name, type, examplelist, memberlist, publist, parlist*,
chlist*, seclist*, abstract)>

<!ELEMENT name (#PCDATA)>
<!ELEMENT type (#PCDATA)>
<!ELEMENT abstract (#PCDATA|cite|dbxref|sub|sup|p|li|i|ol|reaction|pre)*>
<!ELEMENT examplelist (example*)>
<!ELEMENT example (#PCDATA|protein|dbxref|cite)*>
<!ELEMENT publist (publication*)>
<!ELEMENT memberlist (dbxref*)>
<!ELEMENT protein (#PCDATA|protein)*>

<!ATTLIST interpro
  id                  ID    #REQUIRED>
<!ATTLIST dbxref
  db                  CDATA #IMPLIED
  dbkey               CDATA #IMPLIED
  name                CDATA #IMPLIED>
<!ATTLIST protein
  sptr_ac             CDATA #REQUIRED
  status              (?|T|P|F|N) #IMPLIED
  start               CDATA #IMPLIED
  end                 CDATA #IMPLIED>
```

Web Services

- Interfacce per accedere a un servizio per le diverse attività di ricerca e recupero dell'informazione in rete tramite XML
- Consentono l'accesso ai dati in maniera intelligente (comprensione semantica, individuazione dei contenuti) da parte di applicazioni software
- Possono avvalersi di standard per la descrizione dei servizi esistenti (WSDL) e per la loro identificazione (UDDI) e aggregazione (WSFL)
- Necessitano di dati sulle banche dati (metadata)

WSDL: chi fa cosa

Web Services Description Language (WSDL)

- Standard per la descrizione dei Web Services
- Comprende localizzazione, modalità di accesso e descrizione dettagliata
- Funzionalità astratte e dettagli per l'accesso concreti
- WSDL Binding: implementazione di WSDL per SOAP, HTTP, MIME

WSDL: XEMBL.wsdl.xm

```
<definitions name="XEMBL"targetNamespace=http://www.ebi.ac.uk/XEMBL
    xmlns:tns=http://www.ebi.ac.uk/XEMBL
    xmlns:xsd=http://www.w3.org/2001/XMLSchema
    xmlns:soap=http://schemas.xmlsoap.org/wsdl/soap/
    xmlns:soapenc=http://schemas.xmlsoap.org/soap/encoding/
    xmlns="http://schemas.xmlsoap.org/wsdl/
    <documentation>Documentation of this Web Service, together with a
    sample client and links to Bsml (Labbook, Inc.) and AGAVE
    (DoubleTwist, Inc.) can be found at the European Bioinformatics
    Institute http://www.ebi.ac.uk/xembl/</documentation>
+ <message name="getNucSeqRequest" xmlns:tns="http://www.ebi.ac.uk/XEMBL">
+ <message name="getNucSeqResponse">
+ <portType name="XEMBLPortType">
+ <binding name="XEMBLServiceBinding" type="tns:XEMBLPortType"
+ <service name="XEMBLService">
</definitions
```

WSDL: XEMBL.wsdl.xml <service> </service>

```
<service name="XEMBLService">
  <documentation>Returns full information on EMBL Nucleotide Sequences
  formatted as Bsml XML or Agave XML. I.e. returns sequence itself,
  cross-references, taxonomy, literature, full feature information,
  etc.</documentation>
  <port name="XEMBLPort" binding="tns:XEMBLServiceBinding">
   <soap:address
         location="http://www.ebi.ac.uk:80/cgi-bin/xembl/XEMBL-SOAP.pl" />
  </port>
</service>
```

WSDL: XEMBL.wsdl.xml <binding> </binding>

```
<binding name="XEMBLServiceBinding" type="tns:XEMBLPortType">
 <soap:binding style="rpc"
               transport="http://schemas.xmlsoap.org/soap/http" />
 <operation name="getNucSeq">
  <soap:operation soapAction="http://www.ebi.ac.uk/XEMBL#getNucSeq" />
  <input>
    <soap:body use="encoded" namespace=http://www.ebi.ac.uk/XEMBL
          encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" />
  </input>
  <output>
    <soap:body use="encoded" namespace=http://www.ebi.ac.uk/XEMBL
          encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" />
  </output>
 </operation>
</binding>
```

WSDL: XEMBL.wsdl.xml <portType> </portType>

```
<portType name="XEMBLPortType">
  <operation name="getNucSeq">
    <input message="tns:getNucSeqRequest" name="getNucSeq" />
    <output message="tns:getNucSeqResponse" name="getNucSeqResponse" />
  </operation>
</portType>
```

WSDL: XEMBL.wsdl.xml <message></message>

```
<message name="getNucSeqRequest" xmlns:tns="http://www.ebi.ac.uk/XEMBL">
  <part name="format" type="xsd:string">
    <documentation>Input parameter that indicates the result format that
                   should be returned. Legit values: Bsml or sciobj.
                   Defaults to Bsml if format not recognised.
    </documentation>
  </part>
  <part name="ids" type="xsd:string">
    <documentation>A space delimited list of international Nucleotide
                   Sequence accession numbers (IDs).
                   For example: "HSERPG U83300 AC000057".
                   Minimum number of IDs is 1.
    </documentation>
  </part>
</message>

<message name="getNucSeqResponse">
  <part name="result" type="xsd:string">
    <documentation>An XML formatted result in either Bsml or AGAVE
                   format.</documentation>
  </part>
</message>
```

WSFL: comporre servizi complessi

Web Services Flow Language (WSFL)

- Consente di descrivere insiemi di web services
- Flow models: specifica come utilizzare un insieme di web services per raggiungere un certo obiettivo
- Global models: descrive le interazioni tra più web services e il comportamento globale di un insieme
- Recursive composition: ogni flow model o global model viene considerato come un nuovo web service e rientrare in altri modelli

UDDI: la ricerca dei servizi

Universal Description, Discovery and Integration (UDDI)

- Realizzazione di un framework per la descrizione dei Web Services, indipendente da HW e SW
- Consente la creazione di registri di Web Services
- Basato su standard World Wide Web Consortium (W3C) and Internet Engineering Task Force (IETF)
- eXtensible Markup Language (XML), HyperText Transfer Protocol (HTTP) and Domain Name System (DNS) + XML Protocol

Modelli, schemi, ontologie

- Da un record ai record di un'altra banca dati tramite ricerca mediata da interfaccia:
  - Semantica e ambito ben definiti
  - Determinabile automaticamente
    - Ricerca tramite interfaccia standard astratta
    - Esecuzione della richiesta sulla banca dati
    - Restituzione del risultato in formato standardizzato
  - Modelli, schemi, ontologie condivise
- Limitazioni:
  - Diffusione di modelli e strumenti condivisi
  - Competenze informatiche di alto livello

"Middleware is not magic" (C. Goble)

- Qualità dei dati
- Gestione del contenuto dei db (vocabolari controllati)
- Creazione e aggiornamento delle ontologie
- Gestione delle informazioni sull'origine dei dati e la release
- Strumenti appropriati, da usare in maniera appropriata
- Difficoltà nell'accedere all'annotazione a testo libero

Esempi

- SRS (Sequence Retrieval Software)
- CABRI (Common Access to Biological Resources and Information)
- AHMII (Agent to Help Microbial Information Integration)
- GBIF (Global Biodiversity Information Facility)

SRS - Sequence Retrieval Software

- SRS è un esempio di integrazione locale di banche dati eterogenee in maniera semplice ed efficiente
- L'approccio originale di SRS consiste in
  - Banche dati disponibili localmente come "flat file"
  - Definizione di sintassi specifiche per l'estrazione dei dati
  - Utilizzo di link interni espliciti e impliciti tra banche dati
  - L'integrazione trasparente con applicazioni
  - L'integrazione esterna tramite link HTML

Flat files

- I “flat files” sono file di solo testo
  - Possono includere solo il testo, il dato
  - Non possono includere nessun carattere di controllo (formattazione)
  - Non possono includere immagini, altri elementi multimediali, altri contenuti binari
  - Spesso, i caratteri sono limitati al set ASCII base (0 – 127)

Flat files: vantaggi

- I vantaggi derivanti dall'utilizzo di flat files sono
  - Molte informazioni già disponibili non saranno mai strutturate diversamente (80%, in calo)
  - Formato molto diffuso
  - È "leggibile" e adatto anche agli operatori
  - Non necessitano di software costosi
  - Possono includere dati complessi, in maniera articolata, utilizzando un'apposita sintassi
  - Sono facili da indicizzare

Flat files: svantaggi

- Gli svantaggi dell'utilizzo di flat files sono
  - Difficoltà di gestione e aggiornamento delle banche dati
  - Mancanza di controllo di qualità dei dati
  - Mancanza di un linguaggio di interrogazione
  - Scarsa o assente modellizzazione degli oggetti biologici descritti
  - Scarsa o assente strutturazione dei dati

Flat files e DBMS

- DBMS per gestire i dati
  - Database relazionali o a oggetti consentono di gestire in maniera soddisfacente le banche dati
  - Lo sviluppo dei modelli, il controllo di qualità e la gestione dei dati tramite DBMS
- Flat files per scambiare i dati
  - Semplicità nel creare flat files come export, anche con struttura complessa e articolata
  - Conservazione della qualità dei dati
  - Semplicità di trasferimento

Flat files e XML

- I file XML sono flat files
  - Conservano vantaggi flat files
  - Semantica introdotta tramite sintassi DTD
  - È facile analizzare/estrarre dati
- XML per scambiare i dati
  - Conservazione della qualità dei dati
  - Semplicità di trasferimento
  - Vantaggi derivanti da linguaggi collegati

SRS – Dai flat files alle entries

- Flat files per SRS
  - Banche dati in formato flat file/XML
  - Ogni db possiede una sua specifica sintassi, corrispondente alla struttura dati o DTD
  - Analizzando sintatticamente i flat files, SRS è in grado di identificare tutte le informazioni che si riferiscono a un singolo elemento o record
  - Queste costituiscono una entry

```
Strain_number LMG 1(t1)
Other_collection_numbers CCUG 34964;NCIB 12128
Restrictions Biohazard group 1
Organism_type Bacteria
Name Phyllobacterium rubiacearum, (ex Knsel 1962) Knsel 1984 VL
Infrasubspecific_names -
Status Type strain
History <- 1973, D.Knsel
Conditions_for_growth Medium 1, 25C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks Stable colony type isolated from LMG 1. See also Agrobacterium sp.
    LMG 1(t2)

Strain_number LMG 1(t2)
Other_collection_numbers -
Restrictions Either Biohazard group 1 or Biohazard group 2
Organism_type Bacteria
Name Agrobacterium sp.
Infrasubspecific_names -
Status -
Other_names Phyllobacterium rubiacearum, (ex Knsel 1962) Knsel 1984 VL
History <- D.Knsel (Phyllobacterium rubiacearum)
Conditions_for_growth Medium 16, 28C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks One (t2) out of two stable colony types isolated from the original
    culture LMG 1.
```

SRS – Dalle entries ai fields

- L'analisi sintattica dei flat files permette a SRS di identificare i fields all'interno di un entry
- Un Field (campo) è quella parte dell'entry che si riferisce a una particolare informazione
- I Fields possono a loro volta includere subfields, a seconda della complessità della struttura dati e della relativa sintassi
- Elementi DTD possono essere tradotti direttamente in field

Strain_number LMG 1(t1)
Other_collection_numbers CCUG 34964; NCIB 12128
Restrictions Biohazard group 1
Organism_type Bacteria
Name Phyllobacterium rubiacearum, (ex Knsel 1962) Knsel 1984 VL
Infrasubspecific_names -
Status Type strain
History <- 1973, D.Knsel
Conditions_for_growth Medium 1, 25C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks Stable colony type isolated from LMG 1. See also Agrobacterium sp. LMG 1(t2)

SRS – Gli indici

- Qualunque parte della entry può essere indicizzata
  - Un indice speciale viene creato come mezzo d'accesso principale a ciascuna entry
  - Gli indici sono spesso creati sui contenuti dei singoli fields, così che la ricerca possa essere fatta selezionando solo determinati fields
  - Le chiavi degli indici possono comprendere una o più parole, quando queste hanno un significato nel loro insieme (keywords)

SRS – Gli indici

- È possibile non includere negli indici termini aventi un significato particolare
- I numeri e le parole composte da una sola lettera (o un numero limitato di lettere) possono essere esclusi dall'indice
- Alcuni fields possono non essere indicizzati
- È possibile creare indici relativi a più fields

SRS – I links

- I collegamenti (links) tra banche dati possono essere definiti in maniera
  - Esplicita, quando un termine è appositamente inserito in un field come riferimento a una entry di un'altra banca dati
  - Implicita, cercando termini comuni all'interno di fields predefiniti di banche dati diverse

SRS – I links espliciti

- Esplicito riferimento a un'altra banca dati

Other_collection_numbers CCUG 34964; NCIB 12128

Literature DSM ref.no. 72; DSM ref.no. 1300

EMBL: X52289

SRS – I links impliciti

- Termini comuni in banche dati diverse

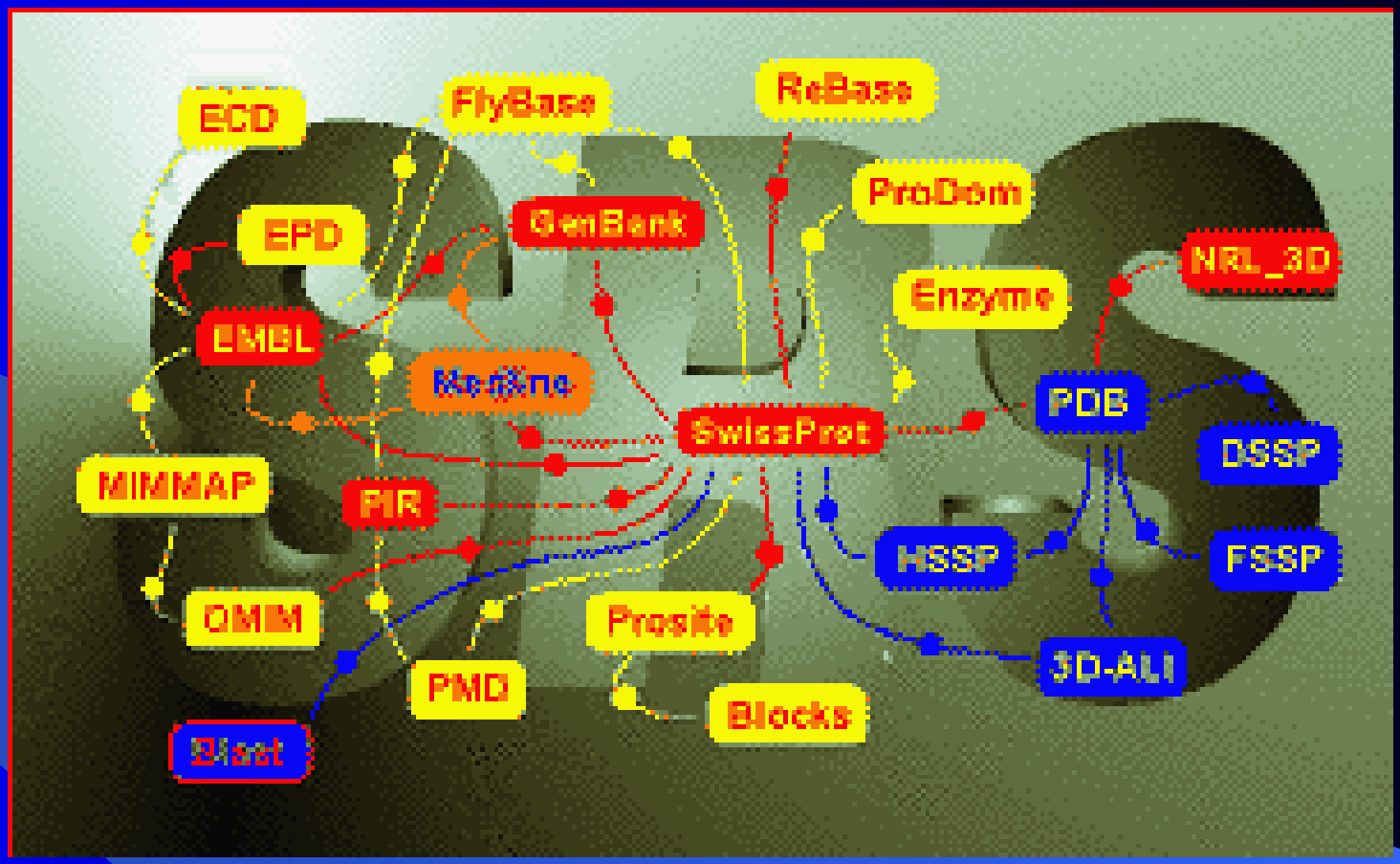
TargetGene: APOE

Constructed_from pMB1, pSC101 and Tn3

Name Gluconacetobacter xylinus subsp. xylinus, (Brown 1886) Yamada, Hoshino and Ishikawa 1998 VL

Literature Nucleic Acids Res 1990;18:4967 [PMID: 2395673]

SRS: mappa dei link



SRS – Possibili estensioni

- SRS è facilmente estendibile
  - Nuove banche dati possono essere aggiunte dando una descrizione della loro sintassi nel linguaggio Icarus o fornendo il DTD
  - È possibile stabilire nuovi collegamenti tra banche dati inserendo xrefs o lasciando che siano identificati da SRS, specificando i fields
  - Nuove releases of banche dati esistenti possono essere facilmente re-indicizzate
  - Molte banche dati sono distribuite in formato “flat file” e con relative sintassi Icarus (qualche DTD)

SRS: operatori link

- SRS consente di utilizzare i link esistenti per le ricerche tramite un apposito operatore: <
  - o swissprot < EMBL
  - o EMBL < swissprot
  - o swissprot < [EMBL-id: X52289]
  - o [EMBL-organism:human] < [medline-pmid:3137981]

SRS – $$$$$$$$$$$$$$

- SRS è (in parte) “free”
  - Dalla versione 6, SRS non è più di pubblico dominio
  - È disponibile solo la versione compilata
  - SRS base è gratuito per gli enti no-profit
  - Le estensioni sono a pagamento

SRS – SRSWWW

- SRSWWW consente l'accesso remoto
  - Il motore di ricerca (WGETZ) viene utilizzato tramite CGI
    - L'utente finale sottopone la propria ricerca tramite form
    - Il server Web richiede l'effettuazione della ricerca a WGETZ passandogli i parametri necessari
    - WGETZ esegue la ricerca e restituisce al server Web i risultati in formato HTML
    - Il server Web restituisce infine i dati all'utente

Gestione di un sito SRS

- Aggiornamento del software
  - Nuove releases (3-4 / anno)
  - Modifiche software / nuove funzioni
- Aggiornamento banche dati
  - Nuove releases (3-4 / anno)
  - Modifica contenuto / struttura file
- Controllo processi
  - Directory temporanee
  - Problemi memoria/disco
  - Analisi degli accessi

Nuove banche dati

- Definizione delle informazioni e analisi delle sorgenti
- Analisi dei link con banche dati esistenti
- Definizione di una struttura dati e di un formato “flat file” o DTD
- Creazione di un’analizzatore di sintassi
- Indicizzazione

Analizzatore sintattico

- Definisce come estrarre i dati dal flat file e come indicizzarli

Applications: Genetic Polymorphism - Hemophilia Diagnosis

Bibliography: Blood 1988;72:1022-1028 [PMID: 3137981]

Literature DSM ref.no. 1026; DSM ref.no. 1300; DSM ref.no. 3394

ComplementaryPrimer: 11.6, 11.1

Conditions_for_growth medium S10, 25C

Analizzatore sintattico

- Linguaggio Icarus, esempio 1

Applications: Genetic Polymorphism - Hemophilia Diagnosis

```
appl:       ~ {$Out $In:[fields c:appl] }
              tag /[^-]+/ {$Uniq:$Ct.trim}
              ( '-' /[^-]+/ {$Uniq:$Ct.trim} )* ~
```

Analizzatore sintattico

- Linguaggio Icarus, esempio 2

Literature DSM ref.no. 1026; DSM ref.no. 1300; DSM ref.no. 3394

```
lit:        ~ {$In:[fields c:lit] $Out $Skip:0}
              word (parola|numero{$Uniq:$Itc})* ~
```

CABRI: obiettivi

Common Access to
Biological Resources and Information

- Distribuzione di materiali biologici di qualità
- Linee Guida per la conservazione del materiale
- Centro Risorse Biologiche virtuale
- Cataloghi integrati tramite SRS
- Integrazione con db esterni
- Shopping cart

CABRI: partners e materiali

Partners:

- BCCM, CABI, CBS, CIP, DSMZ, ICLC, NCCB, NCIMB (collezioni)
- IST, CERDIC (ITC)

Materiali:

- Microrganismi (Batteri, lieviti, funghi filiformi)
- Linee cellulari animali e umane, ibridomi, linee B tip. HLA
- Plasmidi, fagi, virus, sonde DNA
- Complessivamente più di 100.000 risorse

CABRI: struttura dati

Per ogni materiale, identificati:

- Minimum data Set (MDS): dati essenziali, necessari per identificare la risorsa
- Recommeded Data Set (RDS): dati utili per una descrizione precisa della risorsa
- Full Data Set (FDS): tutti I dati disponibili sulla risorsa

Per ogni informazione, linee guida per l'inserimento dei dati:

- Descrizione testuale dettagliata
- Liste di termini e vocabolari di riferimento
- Sintassi predefinite

CABRI: Data sets

Data set	Field label	Catalogues
MDS	Strain_number	All
MDS	Other_collection_numbers	All
MDS	Name	All
RDS	Race	All
MDS	Organism_type	All
MDS	Restrictons	All
MDS	Status	All
MDS	History	All
RDS	Misapplied_names	All
RDS	Substrate	All
RDS	Geographic_origin	All
RDS	Sexual_state	All
RDS	Mutant	All
FDS	Genotype	DSMZ
……	…….	…….

CABRI: Name field

Field	Name
Description	Full scientific and most recent name of the strain. It includes: ▪Genus name and species epithet ▪Subspecies ▪Pathovar ▪Authors of the name ▪Year of valid publication or validation ▪Approbation of the name
Input process	Enter full scientific name as given by depositor and confirmed (or changed) by collection. Names of authors of the name, year of valid publication or validation and approbation are included after a comma. Values for approbation: AL = approved list, c.f.r. IJSB 1980 VL = validation list, in IJSB after 1980 VP = validly published, paper in IJSB after 1980 **Reference list: DSMZ list of bacterial names**
Required for	MDS

CABRI: Reference paper field

Field	Reference paper
Description	Original paper [if available]
Input process	New entries: JournalTitle Year; Volume(issue): beginning page#-ending page# The title is abbreviated following international standard rules (ISSN). Abbreviations are without dot. Authors and title of the article are not mentioned. The reference can be followed by the Pubmed ID enclosed within square brackets as follows: [PMID: 1234567], where '1234567' is the Pubmed ID of the paper
Required for	MDS

CABRI: integrazione

Per ogni catalogo:

- Link HTML a db riferimento (media, hazard, etc…)

Per ogni materiale:

- Link SRS tra cataloghi, basati su dati espliciti e impliciti (Other_collection_numbers)

Per tutti I cataloghi:

- Link HTML basati Pubmed ID verso Medline
- Link SRS / HTML per EMBL Data Library

CABRI: ricerca

CABRI Simple Search:

- Ricerca per ID, nome, tutti I campi restanti
- Ricerca per sinonimi

SRS standard query form:

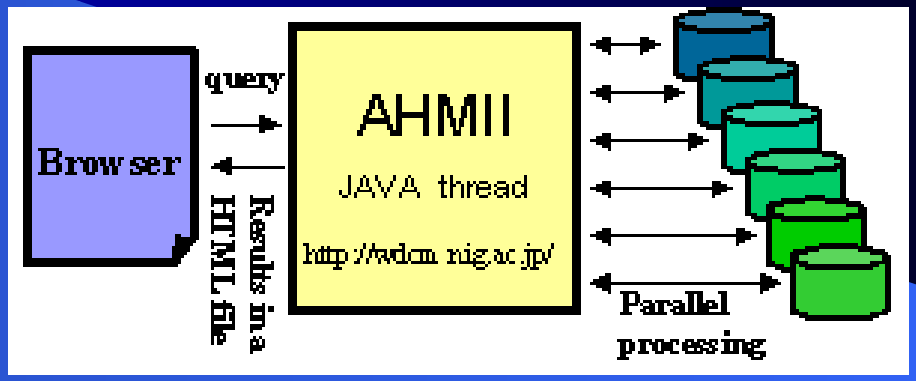
- Utilizzo appieno link SRS
- Gestione viste personalizzate
- Selezione singoli field
- Accesso indici SRS

CABRI HyperCatalogue:

- Indici statici per materiale e per catalogo
- Accesso finale alla descrizione dettagliata tramite SRS

AHMII: Obiettivi

- Accesso a più servizi già esistenti online
- Cataloghi di batteri, funghi e lieviti, linee cellulari
- Liste di nomi scientifici di microorganismi
- Scalabilità

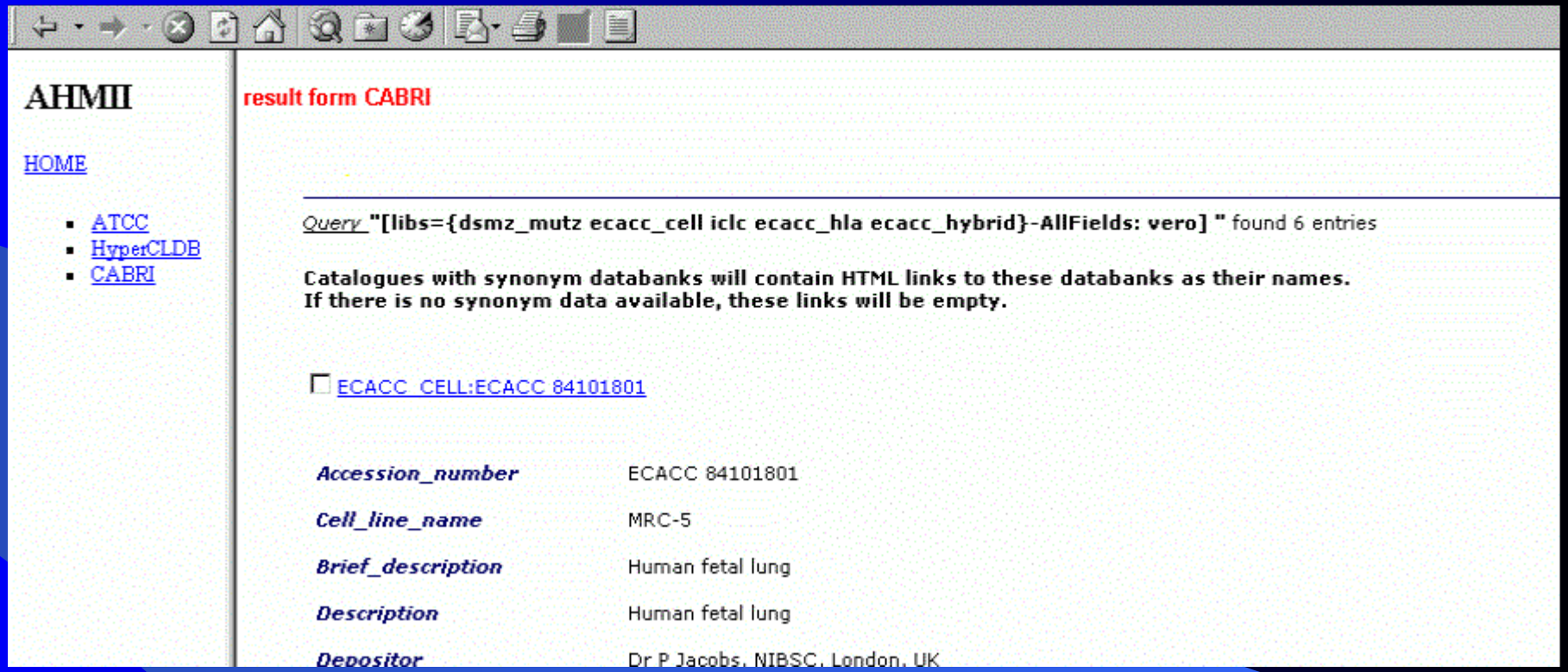


AHMII: Strategia

- Accesso parallelo a più banche dati tramite Internet
- Preparazione e invio di query ad hoc per il sito remoto in emulazione di “form”
- Recupero e visualizzazione del file HTML all’interno della pagina di risposta

AHMII: Query form

AHMII: Results



GBIF: obiettivi

Global Biodiversity Information Facility (GBIF)

- Allestimento rete internazionale integrata (interlinked) banche dati biodiversità
- Segretariato centrale e partecipazione allargata tramite lettera d'intenti a nodi e fornitori dati
- Obiettivo iniziale comprende banche dati su specie e risorse biologiche
- European Network for Biodiversity Information

GBIF: strumenti

XML XML XML

- XML come strumento di riferimento
- WSDL per la descrizione dei Web services
- UDDI per l'allestimento di registri dei Web services
- ABCD (Access to Biological Collection Data) come schema dati di riferimento

- Il resto al prossimo workshop!

Alcune idee prese da…..

- Dagstuhl Seminar 03051:
  Information and Process Integration: A Life Science Perspective,
  http://www.dagstuhl.de/03051/

Con la collaborazione di…..

Idee raccolte e discusse con…

- Carole Goble, University of Manchester, UK
- Thure Etzold, LionBioscience, UK

E con la collaborazione di…

- Domenico Marra
- Beatrice Iannotta
- Assunta Manniello

Istituto Nazionale per la Ricerca sul Cancro, Genova