THOMSON SCIENTIFIC

GENESEQ – Guidelines and policies

Colin Williams – Editorial & Content manager GENESEQ

12th February 2007

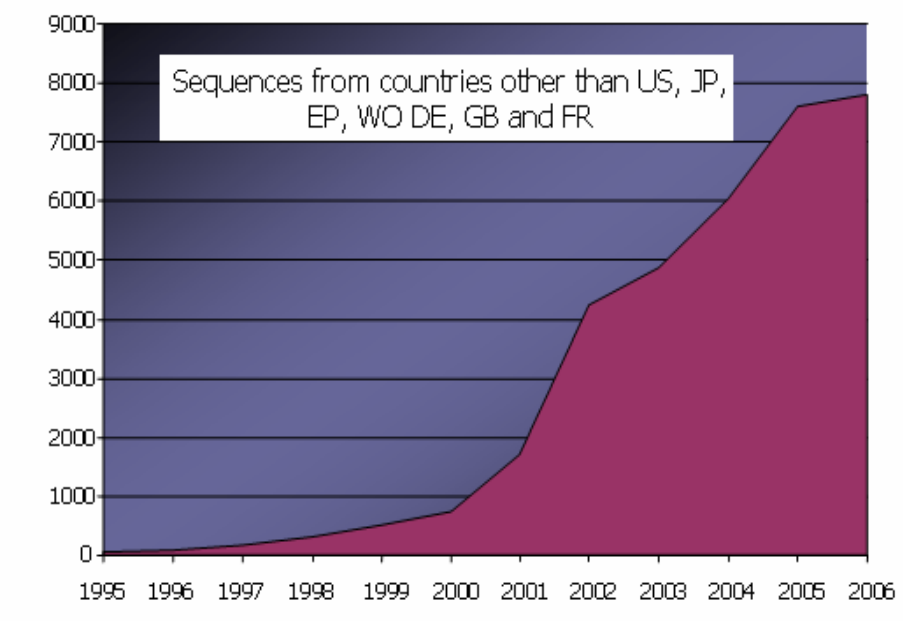
Agenda

- Current database status
- Coverage and selection
- Timeliness
- Indexing policies
  - Organism and keywording
- Quality assurance

Current database status

(Release 200626)

- 2,738,082 protein sequences
- 5,531,614 nucleic acid sequences
- 116,891 unique patents
- 12,590 patents, 275,026 amino acid and 505,130nucleic acid sequences added in 2006
- Increase in activity from SE Asia (especially CN and TW)
- “Mega-patent” indexing – release 24,690,482 sequences from 78 patents over coming months
- Database will increase in size up to ~33million sequences



Coverage and selection

- Cover 41 countries from 1981.
  - Includes Research Disclosures from Kenneth Mason publishing
  - ~20% of records are derived from non-English language patents
- GENESEQ covers only basic patent family members
- Retrospective coverage
  - Novel sequences covered prior to 1992
- Based on basic/equivalent status in the DWPI patent family
  - Provides refined system of relationships defined by the equivalency algorithm complemented by human expert knowledge.
  - Basic status is applied to the first patent of a family received by the Thomson Scientific system
- Identification of sequence containing patents
  - Automated selection of all patents which potentially contain sequence data
  - Patent selection manually screened and tagged for processing
  - Refer to published sequence listings and ensure completeness

Selection policy (CNTD)

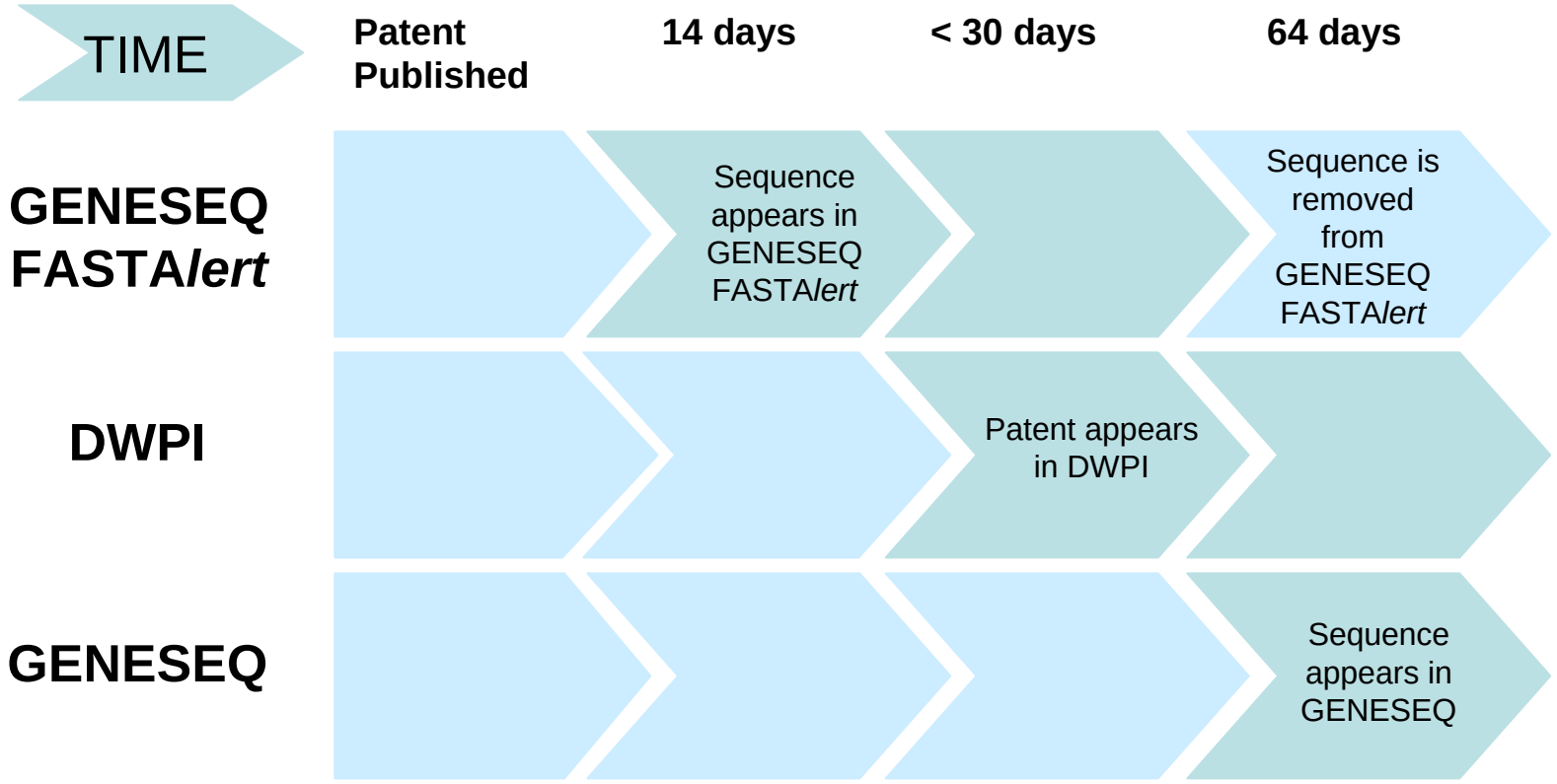
- All sequences which fit selection criteria are indexed regardless of location within the specification

- Sequences are selected and indexed if:
  - Nucleic acid sequences are 10 residues or more or:
  - sequence is 8/9 bases, claimed and within the sequence listings and:
  - The sequence contains 2 bases which do not equal n
  - Amino acid sequences are 4 residues or more and:
  - 2 residues are represented by a letter other than X

- DWPI coverage of sequences outside GENESEQ criteria
  - Any sequence below 4 amino acids or 8 nucleic acids if included in claims.

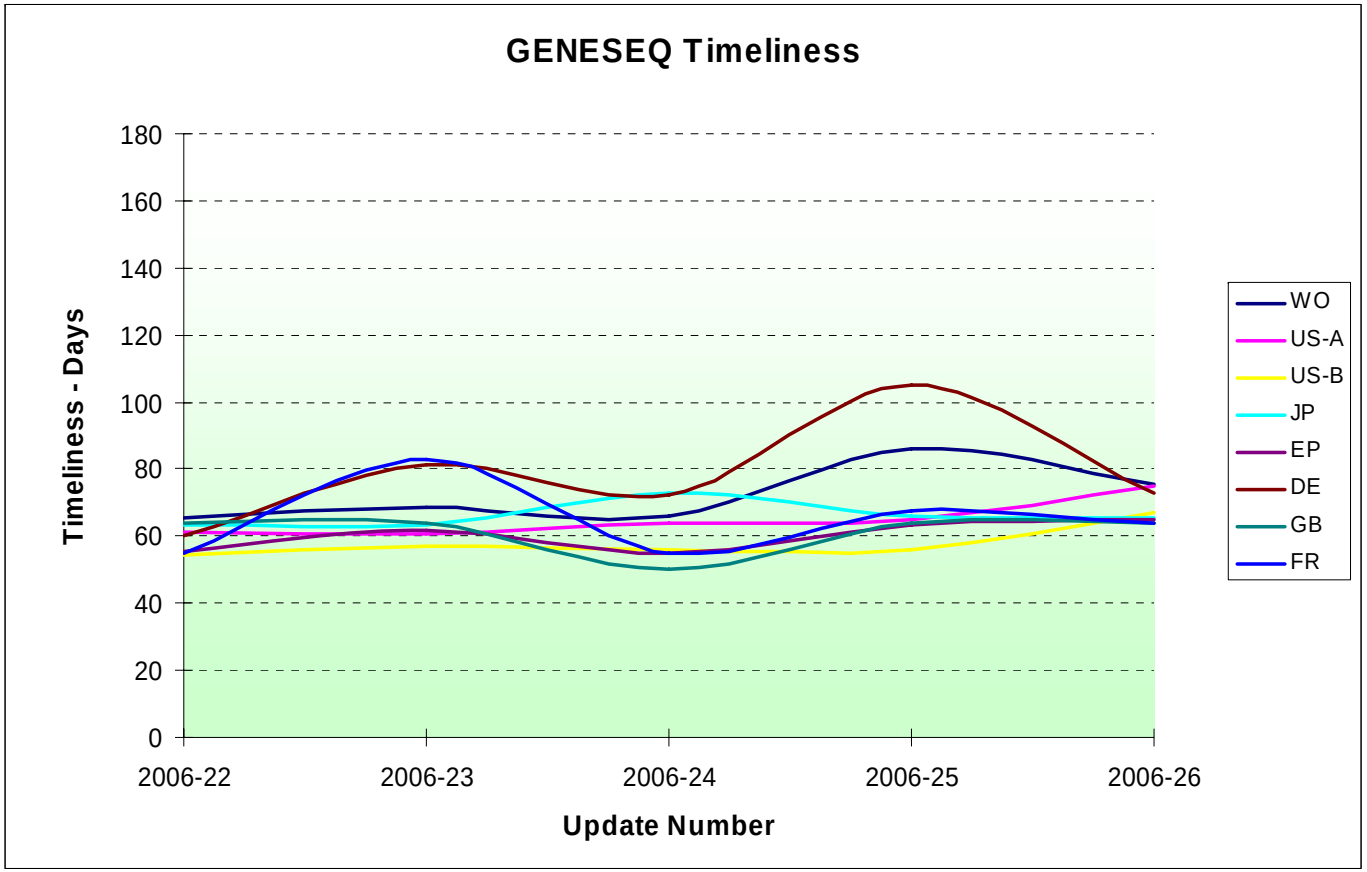
- GENESEQ does not cover sequences referenced in the patent but not explicitly shown, e.g. GenBank accession numbers

DWPI/GENESEQ timeline



Timeliness

- Average timeliness for major GENESEQ countries over last five updates of 2006 are shown below



- **WO 72.6**
- **US-A 65**
- **US-B 58.5**
- **JP 66.5**
- **EP 60**
- **FR 65**
- **DE 79.8**
- **GB 59.3**
- **AVE = 65.8**

Timeliness Snapshot

- In the last incremental product update of GENESEQ in 2006:
  - 82% of patents were at 10 weeks or less after publication date
  - 52% of patents were at 9 weeks or less after publication date

Indexing policies

- American English used since beginning 2004
- Keywording derived from tightly controlled thesaurus.
- Avoid repetition in keywording
- Peptide <40AA if unspecified by patentee
- Feature table, descriptor line

Organism assignment guidelines

- OS line – organism information is applied to the specific sequence when it is explicitly specified in the patent
- NCBI taxonomy browser is reference

  http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy

- Synthetic vs unidentified
  - Synthetic – Artificially produced molecule.
  - Chimeric – Chimeric + organism 1 + organism 2
  - Unidentified – The organism from which the sequence is derived is unknown
  - Synthetic + organism – For non-natural mutated sequences, synthetic analogues of wild type sequences etc

Keywording

- Split into 3 sections:
- Technology focus
  - Preferred term will always be the most specific e.g. DNA vaccine vs Vaccine
  - Methodological terms applied only when directly relevant to the record
- Disease/activity
  - Large numbers of related terms will be grouped into a higher term (with exception to claimed terms) e.g. asthma, bronchitis, pneumonia, emphysema….(n)….Will become respiratory disease.
- Sequence specific
  - Includes protein/gene name
  - Abbreviations are given in full where possible
  - Enzyme names – Derived from in-house thesaurus or http://www.iubmb.unibe.ch/
  - Both patentee and standardised name given

Quality assurance processes

GENESEQ

- Aim: To continually improve consistency in indexing and give maximum searchability for users
  - "Six sigma" quality standard introduced Mid 2006
  - 12 areas of Value add content assessed
  - Statistically identify areas of product content where quality can be improved
  - Manual data capture is double keyed and guaranteed to 99.995% accuracy for good quality images.

Colin.Williams@thomson.com

+44 (0)207 424 2163