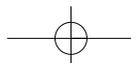
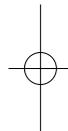
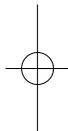




SCIRUS

White Paper

How Scirus Works



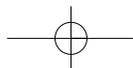
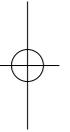
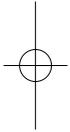
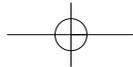




Table of Contents

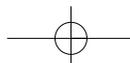
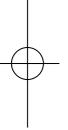
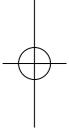
Abstract	
1 Introduction	5
2 Pinpointing Results: The Inverted Pyramid	6
3 Seed List Creation and Maintenance	7
4 Focused Crawling	8
5 Database Load	9
6 Classification	10
Subject Classification	
Information Type Classification	
7 Scirus Index	12
8 Query	13
9 Search	14
Basic Search	
Refine Search	
Advanced Search	
10 Ranking	16
Term Frequency	
Link Analysis	
11 Results	18
12 Conclusion	19
13 References	20
14 Resources	21
15 Glossary	22

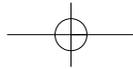


Abstract

This paper answers the question: How does Scirus work? Scirus, a free Web search engine for scientific information, uses sophisticated search technology to create the world's most comprehensive science-specific index. The Scirus Index is comprised of a unique combination ranging from free Web sources to article databases. The process of gathering and classifying the data in the Scirus Index is described. Search functionality, ranking and search refinement are explained in this paper.

Keywords: Science, Technology and Medicine (STM), Web Search Engine, Speciality Search Engine, Focused Crawling, Classification, Scientific Web.





1

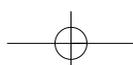
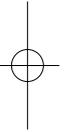
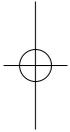
Introduction

Scirus is the most comprehensive science-specific Web search engine available on the Internet. Driven by the latest search engine technology, it enables researchers and students searching for scientific information on the Web to chart and pinpoint the information they need quickly and easily.

Scirus is powered by search technology provided by Fast Search & Transfer™ (FAST™). FAST search technology is currently deployed by a wide range of global customers and partners including AT&T, CareerBuilder, Dell, Freeserve, IBM and Reuters.

Speciality search engines – also called vertical or topical Web search engines – focus on specific subject areas. Elsevier has worked in partnership with FAST to create unique processes that ensure that Scirus is the most comprehensive science-specific index on the market today. Locating scientific information on the Web is fast and efficient with Scirus because it:

- Focuses only on websites containing scientific content and indexes those sites in-depth.
- Searches the world-wide-web for free sources of information such as scientist home pages and university websites.
- Searches the world's largest database of scientific, technical and medical journals.
- Locates pre-print, peer-reviewed articles and patents.
- Provides an intuitive interface and advanced search features that makes it easy to use.
- Provides unique science-specific Web search functionalities, such as searching on author, within specific sources or subject areas

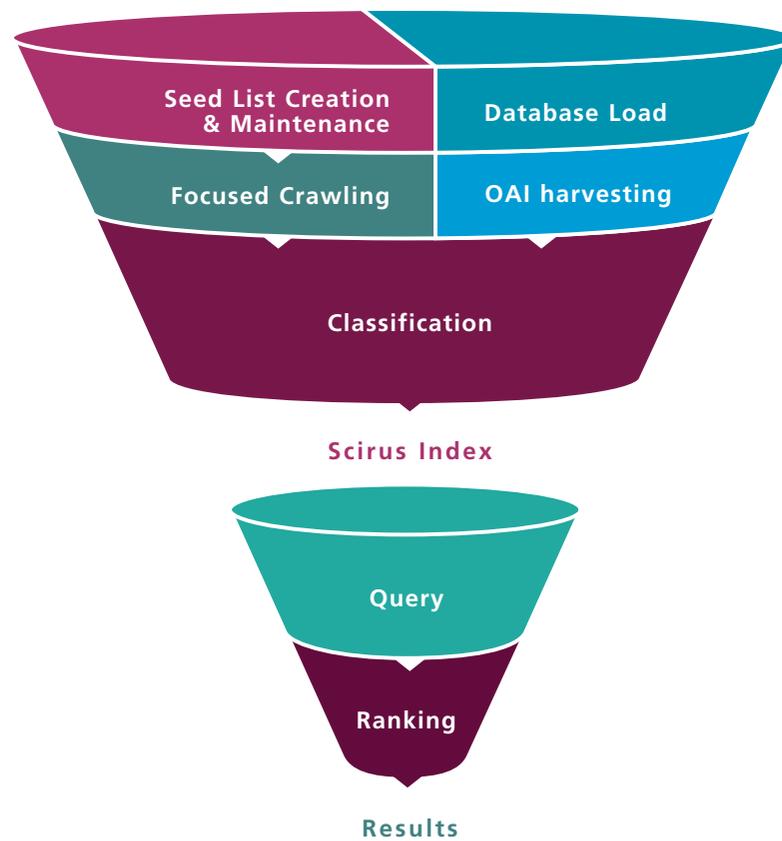


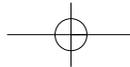


2

Pinpointing Results: The Inverted Pyramid

An efficient search engine is able to pinpoint results quickly and efficiently. An inverted pyramid best depicts the process Scirus uses to pinpoint results. At each phase in the process the data is increasingly refined. As a result, the Scirus Index is a database of relevant, science-specific information.





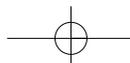
3

Seed List Creation and Maintenance

The seed list is the basis on which Scirus crawls the Internet. The Scirus seed list is created by a number of methods including:

- An automatic URL extractor tool identifies new scientific seeds based on a link analysis of the most popular sites in specific subject areas.
- Elsevier publishing units are periodically asked to supply a list of sites in their subject area.
- Members of the Scirus Scientific, Library and Technical Advisory Boards provide input on an ongoing basis.
- Webmasters and Scirus users regularly submit suggestions for new sites.
- Easily identifiable URLs (such as www.newscientist.com) are added on a regular basis.

Because the seed list only contains URLs that have been manually checked for scientific content Scirus is able to crawl the Internet in an efficient, focused way ensuring deeper and fuller coverage of scientific web sites.





4

Focused Crawling

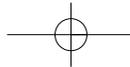
Scirus uses a robot – also known as spiders or crawlers – to “read” the text on the sites found on the seed list.

Unlike general search engines, the Scirus robot doesn’t follow links unless those domains are also on the seed list. This type of focused crawling ensures that only scientific content is indexed. For instance, if Scirus crawls www.newscientist.com it will only read pages that fall under that domain. It doesn’t crawl a link to www.google.com because that URL isn’t on the seed list.

The Scirus robot crawls the Web to find new documents and update existing documents. The process the robot follows is relatively simple:

- A scheduler coordinates the crawl. The job of the scheduler is to prioritise documents for crawling, track the rules set by webmasters in their robots.txt files and limit the number of requests a robot sends to a server.
- Independent machine nodes – sometimes called a “crawler farm” – crawl the Web. They work in tandem and share link and meta information. Each node in the farm is assigned a segment of the Web to crawl.
- The robot collects documents and sends them to the Index. It also stores a copy of the page so that Scirus can show the portion of the document that actually contains the search query term (also called a dynamic teaser or keyword in context).

Robots face a number of challenges when crawling the Web due, in large part, to its exponential growth. Links are added and removed with great frequency. As a result, it is necessary to index both the page and their relationships.



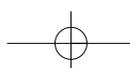
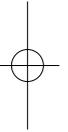
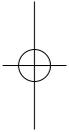
Database Load and OAI Harvesting

While the robot crawls the seed list, Scirus loads data from science-specific sources. The loaded data consists of both partnership and Open Archive Initiative (OAI) sources. Partnership sources currently include ScienceDirect, MEDLINE, BioMed Central, and the US Patent Office.

OAI sources currently include arXiv.org, NASA (incl. NACA and LTRS), CogPrints, Project Euclid (Cornell University) and several Preprint Servers.

The Scirus database will continue to expand to ensure that as many science-specific Web sources as possible are included in the index.

For a complete overview of the Scirus content, visit www.scirus.com





6

Classification

The robot gathers all the pages and puts them into a “working” index. Scirus reads every word that appears on the site and examines where the word appears on the site (title, URL, text). Once the seed list has been crawled and the database has been loaded Scirus is ready to classify the data. The classification process improves the retrieval of science-specific pages and allows the user to perform searches that are targeted towards specific scientific domains or document types.

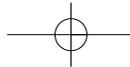
Scirus performs document classification following two different schemes:

- The subject classification identifies scientific domain descriptors that can be assigned to a document. Currently there are 20 subject areas available for selection – such as Medicine, Physics or Sociology – covering all major fields of science.
- The information type classification assigns a document type such as scientist homepage or scientific article. This narrows the search to specific kinds of documents and prevents the retrieval of unwanted pages.

Subject Classification

Scirus maintains a customized linguistic knowledge base for each subject area. The vocabulary on the pages is mapped against the terms from dictionaries which have been compiled through training on a very large, manually pre-classified corpus of scientific texts. The dictionaries are supplemented with entries from domain-specific terminological databases.

The vector terms are weighted single word and multiword expressions. The weight of the classification terms is determined by examining statistical properties from the training corpus – such as the classification strength – and through partial manual maintenance.



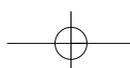
In addition to their classification task, the dictionaries are also used to determine and normalise terms for providing the main keywords relevant for the document. Document meta information, such as the URL of the page and anchor text pointing to the document is also used to refine and improve the subject classification.

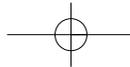
The algorithm for subject classification allows the assignment of multiple categories to a single document, because of the considerable overlap between neighbouring scientific disciplines (such as Neuroscience and Medicine or Psychology and Social Sciences).

Information Type Classification

Scirus uses custom software to analyze the profile of a page and classify the information type. Types that are recognized include scientific abstracts, full text scientific articles, scientist home pages, conference announcements and other page types that are relevant to the scientific domain. The classification algorithm analyses the structure and the vocabulary of a page to assign one of the categories. For instance, scientist home pages can be recognised by looking at structural information – such as the presence of address information, biographical data layout, publication lists – and by the presence of keywords like “homepage”, “publication list” etc.

The structural analysis also allows the extraction of certain information chunks from the analysed pages. In the case of a scientist homepage the module will attempt to extract information like the name and affiliation of the owner of the page and add it to the document attributes.





7

Scirus Index

Once the classification is complete the Scirus Index is ready for searching. Like the Internet, the Scirus Index has grown substantially since it was launched in 2001.

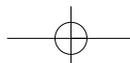
Scirus strives to be the most comprehensive search engine for scientific information on the Web and is constantly growing its index by adding new web sites and sources.

The index of Scirus returns results from the whole Web, including access-controlled sites that other search engines don't index. Scirus currently covers the following information from the Web:

- University pages (domains ending with .edu, .ac.uk and educational sites within other countries)
- Scientific organizations (.org) and conference pages
- Company pages with scientific R&D information or information relevant to researchers (.com)
- Government pages dealing with science, including health and sciences like law (.gov)
- Scientist and author homepages

In addition to Web pages, Scirus indexes a number of journal sources, preprint servers, patent sources, repositories and databases.

We ensure that proprietary content always contains a free layer of information, such as a journal abstract, accessible by our users.



Query

Scirus improves the ranking and relevance of results by implementing “intelligent query rewrites” which are designed to automatically understand the intention of the user and enable more intelligent searching by rewriting the queries. Query transformations performed on the searchers’ behalf include the addition of quotes around common phrases that are detected from the Scirus phrase dictionary and the removal of non-essential search words in the query such as “what is” and “where can I find information about”. Searchers have the option of running a query without the re-writing function.

The screenshot displays the Scirus search interface. At the top, the Scirus logo is followed by the tagline "for scientific information only" and a link to "Latest Scientific News - from New Scientist". A navigation bar includes links for "About Us", "Newsroom", "Advisory Board", "Submit Web Site", "Search Tips", and "Contact Us".

The search interface features a "Basic Search" section with a search box containing the query "optimal number of hidden nodes in neural networks". Below the search box are checkboxes for "All journal sources" (checked), "All Web sources" (checked), and "Exact phrase" (unchecked). A "Search" button is positioned to the right of the search box. Links for "Advanced Search" and "Search Preferences" are also visible.

Below the search box, a summary table shows the search results:

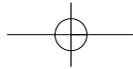
Searched for:	All of the words optimal number of hidden nodes in neural networks
Found:	246 total 92 journal results 154 Web results
Sort by:	relevance date

Below the summary table, there are two buttons: "Save checked results" and "Email checked results". The search results are listed as follows:

- 1. Determining the Optimal Number of Hidden Nodes in a Feed-Forward Neural Network (ResearchIndex)**
Jan 2004
...Determining the **Optimal Number of Hidden Nodes in a Feed-Forward Neural Network**...Artificial **Neural Networks** on Large...Generalization Of A Constructive Algorithm **In** Pattern...Determining the **Optimal Number of Hidden Nodes in a Feed-Forward Neural Network**...**1 Neural Networks** (context... [more hits from](#) [http://citeseer.nj.nec.com/371325.html] [similar results](#)
- 2. Neural networks in geophysical applications**
Mar 2002
...MLP) and **neural networks** with localized...a single **hidden** layer are...the weight of link k to...nh1 the **number of nodes in the hidden**...several output **nodes** and more **hidden** layers. Likewise, RBF **networks** with a single...the center of the localized...function **in** hid- den...the total **number** (ntot) of...behavior of the **neural networks**...Damage OBS **Optimal** Brain Surgeon...a single **hidden** layer are considered. **In** addition...defines the **number** of training...applying **neural networks** to recognition... [http://earth.leeds.ac.uk/~mvdbaan/NN.pdf] [similar results](#)
- 3. Appears in Australian Conference on Neural Networks, ACNN 96, Edited by Peter Bartlett, Anthony Burkitt, and Robert Williamson, ...**
Jun 1999
...**2 Neural Networks** It has been...that an MLP **neural** network, with a single **hidden** layer, can...any degree of accuracy...sufficient **number of hidden**...However, **in** practice, the **number of hidden** layer neurons...impractically large. **In** addition...surface. The **networks** suffer from...requires for 2 **nodes** to be visited...each node...high

On the right side of the results, there is a box explaining the query rewrite: "Your query was rewritten to: 'optimal number' of 'hidden nodes' in 'neural networks'". It notes that quotes were added to common phrases and non-essential words were removed, with a link to "Repeat without rewrite".

Below the query rewrite box, there is a section titled "Refine your search using these keywords found in the results:" which lists various related terms such as "activation functions", "artificial neural network", "back propagation", "bayesian", "cross validation", "generalization", "generalization error", "ieee trans", "nonparametric", "output layer", "perceptron", "radial basis function", and "rule extraction".



9

Search

Scirus has a wide range of features to help users pinpoint the information they're looking for.

Basic Search

The basic search function enables users to specify:

- Search only on the exact phrase.
- Show results from all sources or select either journal or web sources.

Refine Search

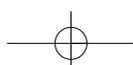
Users can refine their search by selecting from a list of relevant classification terms. These terms are identified by analysing the top 100 results and tabulating the most common classification terms assigned to them.

The "refine your search" term functionality is based on the classification terms added to the document during indexing.

Advanced Search

The Advanced Search option allows the users to customize their search in the following ways:

- Select from a range of 20 searchable subject areas spanning health, life, physical and social sciences.
- Locate data within a specified date range.
- Search by information type – such as scientific conferences, abstracts and patents.
- Search within specific information sources such as journals on BioMed Central or a web source such as NASA.
- Search by journal title, article title or author name.



Advanced Search [Basic Search](#) [Advanced Search](#) [Search Preferences](#)

optimal number of hidden nodes in neural networks The complete document All of the words

AND

_____ The complete document All of the words

Search

Published between	1920 and 2005	
Information types	<input checked="" type="checkbox"/> All <input type="checkbox"/> Abstracts <input type="checkbox"/> Articles <input type="checkbox"/> Books <input type="checkbox"/> Company homepages	<input type="checkbox"/> Conferences <input type="checkbox"/> Patents <input type="checkbox"/> Preprints <input type="checkbox"/> Scientist homepages
File formats	<input checked="" type="checkbox"/> All <input type="checkbox"/> HTML	<input type="checkbox"/> PDF
Content sources	<input checked="" type="checkbox"/> All Journal sources <input type="checkbox"/> BioMed Central <input type="checkbox"/> MEDLINE <input type="checkbox"/> ScienceDirect <input type="checkbox"/> Society for Industrial & App. Mathematics	<input checked="" type="checkbox"/> All Web sources <input type="checkbox"/> CogPrints <input type="checkbox"/> E-Print ArXiv <input type="checkbox"/> Computer Science Preprint Server <input type="checkbox"/> Mathematics Preprint Server <input type="checkbox"/> NASA <input type="checkbox"/> US Patent Office <input type="checkbox"/> Other
Subject areas	<input checked="" type="checkbox"/> All <input type="checkbox"/> Agricultural and Biological Sciences <input type="checkbox"/> Astronomy <input type="checkbox"/> Chemistry and Chemical Engineering <input type="checkbox"/> Computer Science <input type="checkbox"/> Earth and Planetary Sciences <input type="checkbox"/> Economics, Business and Management <input type="checkbox"/> Engineering, Energy and Technology <input type="checkbox"/> Environmental Sciences <input type="checkbox"/> Languages and Linguistics <input type="checkbox"/> Law	<input type="checkbox"/> Life Sciences <input type="checkbox"/> Materials Science <input type="checkbox"/> Mathematics <input type="checkbox"/> Medicine <input type="checkbox"/> Neuroscience <input type="checkbox"/> Pharmacology <input type="checkbox"/> Physics <input type="checkbox"/> Psychology <input type="checkbox"/> Social and Behavioral Sciences <input type="checkbox"/> Sociology

Search



10

Ranking

Scirus uses an algorithm to rank the documents resulting from a query. Algorithms are procedures, or formulas, used to solve a problem. Ranking is based on two basic values: term and links.

Term Frequency

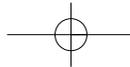
For term value, the location and frequency of occurrence of the terms within the document are measured. The global frequency of the term within the whole index is also taken into consideration. Scirus asks the following questions when looking at term location and frequency:

- Is the term in the title?
- Is the term in the text in a link?
- Where is the term located in the text (top, bottom)?
- How many times is the term used?

To ensure that full-text articles are not ranked higher than title/abstract pages Scirus counts the number of keywords and then divides them by the total number of terms in the document. Scirus also examines the length of the URL.

Short URLs (such as www.microsoft.com) are more relevant than longer URLs (such as www.microsoft.com/help). Scirus does not use meta tags because they are subjective to ranking tweaking by users.

When the terms in a query occur near to each other within a document it is more likely that the document is relevant to the query than if the terms occur at a greater distance. Therefore the proximity of the search terms influence the Scirus ranking.



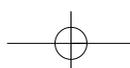
Link Analysis

Scirus uses link analysis as part of its relevancy ranking system. For link value, the number of links to a page is analysed. The cardinality or importance of a page is determined by calculating the number of links to a page. The more links to the page, the higher the ranking. Scirus also analyses the anchor text – the text of a link or hyperlink – to determine the relevance of a site.

Because pages that are indexed via database loading are not crawled and are usually less interlinked, it is not possible to conduct a proper link analysis. These pages are assigned a static score. Every time a new Scirus Index is created the static score is examined for relevance.

Scirus uses a special general terms dictionary with “select” scientific terms to identify which pages deserve a science flag. In cooperation with the Computational Linguistics Department at the University of Munich, the Scirus development team created a large scientific dictionary with scientific terms that occur uniformly across all subject areas. In order to receive a science flag the page needs to meet a threshold number of matches.

The results with the highest rank score are listed first. Each document that matches a query is given a rank value. Static ranking is assigned based on an analysis of the document alone. Dynamic ranking is based on the location of the query terms in the document.



11

Results

To ensure that results analysis is an efficient process, Scirus presents results in a number of innovative ways:

- It “collapses” the site to prevent returning multiple pages of the same Website. Although the content is different, pages from the same domain often look alike. If the user clicks “more hits from” at the end of the citation Scirus will display more matching results from the same Website.
- Dynamic teasers (also called relevant text) return the part of the result relevant to the query and highlight the search terms. For example, if you search for DNA repair the first result has the following teaser:

1. [DRIG - What is DNA repair?](#)
 Apr 2003
 What is **DNA repair**? Figure 1 -**DNA Repair** functions As a major defense against environmental damage to cells **DNA repair** is present in all organisms examined including bacteria, yeast...
[more hits from](#) [http://www.nih.gov/sigs/dna-rep/whatis.html]
[similar results](#)

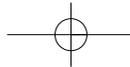
- Sources are branded so that it is clear whether the results are from the Web or loaded databases. For example, BioMed Central results are displayed as follows:

Save checked results
Email checked results

1. [Phylogenomic identification of five new human homologs of the DNA repair enzyme AlkB](#)
Kurowski, Michal A. / Bhagwat, Ashok S. / Papaj, Grzegorz / Bujnicki, Janusz M., BMC Genomics, Dec 2003
 ...is related to known protein structures) Aravind and Koonin found that the Escherichia coli AlkB protein implicated in **DNA repair** is also a member of the 2OG-Fe²⁺ oxygenase superfamily 3 . They proposed that AlkB may catalyze oxidative demethylation...

Full text article available from  **BioMed Central**

[similar results](#)

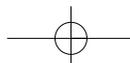
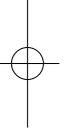
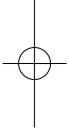
**12**

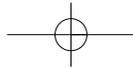
Conclusion

The vast amount of information available on the Internet can make searching a long, complicated process. Speciality search engines like Scirus provide a more productive search by:

- Focusing only on sites with subject-specific data.
- Searching the “deep” web.
- Filtering out irrelevant data.

Scirus, like the Internet itself, is constantly expanding and developing. We work in partnership with FAST to keep pace with evolving technology so that our users can tap into the vast pool of scientific information available on the Internet. Through an in-depth process of filtering information, classification and ranking Scirus' users are able to find scientific information on the Web quickly and easily.





13

References

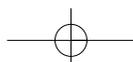
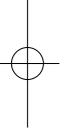
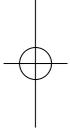
Paper on Text classification for Scirus by ELIXIR (a subsidiary of Fast Search and Transfer ASA) and CIS, University of Munich, year 2001

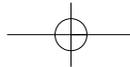
Search Engine Terms as compiled by members of the I-Search Digest.
http://www.cadenza.org/search_engine_terms/index.htm

How a Search Engine Works, SEARCHER, The Magazine for Database Professionals,
May 2001

Search Engine Watch: <http://www.searchenginewatch.com>

FAST technical documentation



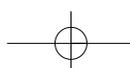
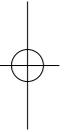
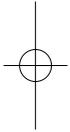


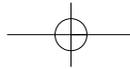
Resources

FAST Search & Transfer (<http://www.fastsearch.com/>)

Open Archives Initiative (<http://www.openarchives.org/>)

Search Engine Watch (<http://www.searchenginewatch.com/>)





15

Glossary

Algorithm

A procedure, or formula, for solving a problem.

Anchor Text

The text contained in (and sometimes near) a hyperlink.

Crawler

The part of a search engine that surfs the web, storing the URLs and indexing the keywords and text of each page.

Deep Web

Content that general search engines can't access. Also known as the Invisible Web.

Dynamic Teaser

Search terms that are highlighted in the search results. Also known as keywords in context.

HTML

Hypertext Mark-up Language is a set of tags used to structure text and multimedia documents posted on the Internet.

Index

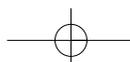
An index – also known as database – is a collection of web pages maintained by a search engine.

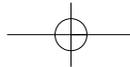
Invisible Web

Content that general search engines can't access. Also known as the Deep Web.

Keyword

A word, a phrase or a group of words, sometimes combined with other syntax used in a query.





Meta tag

A construct placed in the HTML header of a web page, providing information that is not visible to browsers. The most common meta tags are keywords and description.

Nodes

A terminal in a computer network.

PDF

A file format used to represent documents independent of the original application software, hardware, and operating system used to create those documents.

Query

Instructions given to a search engine in order to locate web pages.

Ranking

The process of ordering web sites or web pages by a search engine so that the most relevant sites appear first in the search results.

Robot

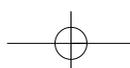
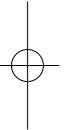
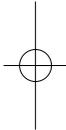
A program that follows hypertext links and accesses web pages. Also known as a Spider or Crawler.

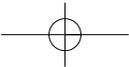
Robots.txt

A text file stored in the top level directory of a web site to deny access by robots to certain pages or sub-directories of the site.

URL

A Universal Resource Locator is a unique address that specifies an Internet resource.





© 2004 Elsevier. All rights reserved. SCIRUS is a registered trademark of Elsevier in the United States and/or other jurisdictions.

Contact Details Scirus®

Radarweg 29
1043 NX Amsterdam
Netherlands
Email address: feedback@scirus.com
For more information, visit: www.scirus.com

