# If INEX is the Answer, what is the Question?

Richard A. O'Keefe

CS Dept, University of Otago

**Abstract.** The INEX query languages allow the extraction of fragments from selected documents. This power is not much used in INEX queries. The paper suggests reasons why, and considers which kind of document collection this feature might be useful for.

## 1 What is the INEX Answer?

INEX [1–3] is all about extracting information from XML document collections. We can distinguish four kinds of IR-like query for semi-structured data:

CO Content Only—a classical information retrieval query to select a document from a collection of documents based on the occurrence of terms and phrases anywhere within it. Example: "find all documents mentioning 'Malacostraca' and 'Edgar Allen Poe'."

CC Content-in-Context—a combination of contexts (paths) and CO queries to apply in those contexts, used to select documents from a collection of documents. Queries like this have been around almost as there have been SGML collections to search in. Example: "find all documents where author mentions 'Edgar Allen Poe' and body mentions 'Malacostraca'."

EC Element-in-Context—a CC-like query is used to select elements from documents in a collection, with each element being treated as if it were a document and reported separately. These are NEXI [4, 5] "Basic CAS" queries. You can see CC queries as BCAS queries that just happen to select `article` elements, but the distinction between CC and EC is useful. Example: "find ⟨bibitem⟩s mentioning 'INEX'."

2S Two-Stage—An EC query is used to select elements, and then a further EC query is used to select portions of those elements. This is not used for highlighting within documents; the elements selected in the second stage are reported separately. Example: "find ⟨back-matter⟩ where any ⟨author-bio⟩ mentions 'Edinburgh' then report contained ⟨bibitems⟩ mentioning 'DAI'."

The INEX Answer is "EC and 2S queries", or "element extraction".

Because an XPath query that is purportedly about some element can examine remote descendants of ancestors of that element, it can be difficult to tell the difference between EC and 2S queries. I regard a query that examines an element and its descendants and at most the attributes of its ancestors as an EC query, others as disguised 2S queries. This classification is sensitive to whether publication date, for example, is an attribute or an element, which is why both EC and 2S queries must be allowed.

## 2 What is Problematic about the INEX Answer?

It turns out that INEX participants have found it very hard to formulate non-trivial EC and 2S queries, and even harder to evaluate them. The INEX'03 [2] topics included thirty Content and Structure queries:

| count | type | tag returned | what that tag means |
|---|---|---|---|
| 14 | CC | `article` | whole articles |
| 3 | EC | `sec` | sections |
| 1 | EC | `abs` | abstracts |
| 1 | EC | `p` | paragraphs |
| 1 | EC | `vt` | *curricula vitæ* |
| 6 | 2S | `sec` | sections |
| 2 | 2S | `abs` | abstracts |
| 1 | 2S | `bb` | bibliography items |
| 2 | 2S | `*` | IR engine's choice |

That is, nearly half of the queries did not exploit the INEX Answer.

One reason for this is simply that there is not a lot of structure that one can usefully exploit in the INEX collection. Basically, there are front matter, including authors, title, and abstracts, body with a whole bunch of variously tagged sections and subsections, and back matter with bibliography and author biographies.

Things changed in INEX'04 [3], but not much. There were 35 CAS topics.

| count | type | tag returned | what that tag means |
|---|---|---|---|
| 8 | CC | `article` | whole articles |
| 2 | EC | `sec` | sections |
| 1 | EC | `abs` | abstracts |
| 1 | EC | `p` | paragraphs |
| 1 | EC | `vt` | *curricula vitæ* |
| 1 | EC | `bib` | entire bibliographies |
| 1 | EC | (`p`\|`fgc`) | paragraphs or figure captions |
| 8 | 2S | `sec` | sections |
| 1 | 2S | `abs` | abstracts |
| 1 | 2S | `bb` | bibliography items |
| 1 | 2S | `p` | paragraphs |
| 1 | 2S | `fig` | figures |
| 1 | 2S | `bdy` | whole bodies |
| 2 | 2S | `*` | IR engine's choice |

A little over three quarters of the INEX'04 CAS queries did exploit the INEX Answer, but how usefully?

Some of these queries are thought-provoking.

- In query 161, the containing `article` must be about access methods for spatial data and text, while the selected `bb` elements need not be about either. They could be about access methods for time series, for example.

– In query 158, the containing `article` must be about the Turing test, while the selected `bdy` element must be about the "turning" test. Nor is it clear why it is useful to see an article without its title, authors, or abstract.
– Query 158 also makes one wonder how a query of the form `about(.//fm,` $x$`) or about(.//abs,` $x$`)` differs from a simple `about(.//fm,` $x$`)`, since `abs` only occurs inside `fm`.
– Query 127 with its `(p|fgc)` reminds us that while the average `p` in the INEX collection has about 300 characters of text, the average `fgc` has about 150 characters. So perhaps more (all?) queries that accept `p` elements should also accept `fgc` elements.
– Query 136, selecting entire bibliographies on the basis of "text" and "categorisation" appearing somewhere, and "SVM" and "Support Vector Machines" appearing somewhere else, reminds us that titles are not a reliable guide to relevance. Who would dream from the title alone that *Bananas in Space* was about "functional programming" using the "Bird-Meertens" formalism?
– Query 142, of the form `//abs[about(...)]`, makes one wonder why it is useful to find an interesting abstract if you cannot tell which article it is an abstract *of*.

Queries must not only be formulated, they must be evaluated. And to evaluate the relevance of an element, you may need a greater or lesser amount of context. As IR researchers well know, words are ambiguous. If you see "Algol is very old", is that talking about the star or the programming language (and if so, which)? If you see "The tables were too crowded", is this a complaint about a paper or a dining hall?

This points out a serious methodological problem in the INEX evaluation procedure. Judges rate elements within the scope of complete articles (which they can and do look at), while users would presumably just see the elements. That is, for CO and CC queries, the judge and the user have the same information available to them, while for EC and 2S queries, the judge has far more information at his or her disposal in making relevance judgements than someone just receiving the paragraphs or sections in question would. For abstracts and sections, this may not be too much of a problem, but paragraph, title, and bibliography item it is almost certainly a distortion. Even for sections, I know that I found myself either able to dismiss an entire article quickly (having looked at a portion that was not part of the selected response) or else having to read the entire article with care to decide what the flagged elements actually *meant* before I could decide how relevant they were. Does it even make sense to talk about a small element *having* any relevance without its context?

## 3 What Might the Question Be?

What should our collection be like for the INEX Answer (element extraction) to be useful?

### 3.1 Strong semantics for markup

Some markup in the INEX collection has strong semantics. An `ead` element should be an e-mail address, nothing else. The `mo`, `day`, and `yr` elements are parts of dates. A `bb` element is always a bibliographic reference. The `abs`, `bb`, and `vt` elements are clearly useful in queries.

Some markup in the INEX collection has presentation semantics. The `it` and `rm` elements select italic and roman faces, but say nothing about why. It is not accidental that none of the queries mention these elements, and it is only regrettable that the evaluation system requires people to judge these elements.

Some markup is structural, without having much semantics. There is nothing to mark the rhetorical structure of a document or the rhetorical force of any element. There is, for example, no distinction between "quoted in support" and "quoted for rebuttal". Structural elements are surprisingly popular in queries, principally `sec` with some `p`. One feels that this may be an artefact of the INEX setup: people are under pressure to select *something* to show that the INEX Answer is useful, and `sec` is the smallest nearly-self-contained element. It is difficult to imagine any queries where `ss1` or `ss2` would be meaningful choices.

An INEX Question really needs a wider range of elements with strong semantics: `exercise`, `example`, `poetry` (in the INEX DTD, but apparently not used anywhere), `warning`, `listing`, `scene`, `design.pattern`, that kind of thing.

### 3.2 Memorable markup

You cannot ask about tags that you cannot remember. A DTD or Schema may contain more tags than people can recall; the present 192 is almost certainly too many. Tag names may be difficult to recall. The present DTD uses names that have been heavily abbreviated, like ⟨ilrj⟩. Users may not be provided with enough information about the meaning of tags; how is an ⟨ilrj⟩ different from other paragraphs?

This suggests that the markup assumed in queries should contain not too many tags, which should not be too heavily abbreviated, and should be clearly explained to query users.

The "query DTD" need not be the actual DTD used for markup. This is already the case in INEX, where several kinds of paragraph are mapped to ⟨p⟩. Architectural form processing (a major concept in SGML) means that a small "authoring" DTD can be mapped to a rich one and that a rich DTD can be mapped to a small "querying" DTD.

### 3.3 Low coupling

What really matters is not how big the fragments are but how tightly they are coupled to their context. The Wall Street Journal documents from TREC are smaller than most of the IEEE `sec` elements, but they were written to be free-standing. The `bb` and `vt` elements make good sense as fragments in the existing INEX collection because they depend hardly at all on their context. Abstracts

are crafted to be fairly self-contained. In contrast, `p` elements are so tightly linked to their context as to be difficult to judge, even though they are bigger than most `bb` elements. The very smallest body extracts that work are `sec`, and even they depend too much on context for comfort.

We need a collection of documents which have pieces whose relevance can be judged on their own.

### 3.4 Some coupling

If the fragments we want are not coupled to their containing document at all, why are they not stored as free-standing documents in the first place? There has to be enough coupling so that the first EC filter usefully limits the scope of the second EC filter.

### 3.5 Sizeable fragments

If you find a relevant `sec`, do you not want to know what article it came from in case there is more good stuff there, or to find the author's address to write for more information? One reason you might not want to do this is if the "documents" are too big to examine or or too unlikely to contain other relevant material.

### 3.6 Examples

2S From the Otago Daily Times, issues in 2003, find stories about Don Brash.
Newspapers contain many stories with low or no coupling. This is almost a WSJ query. The trick is to find queries with more constraints on the container (issue).

2S From the Otago Daily Times, issues since 2000 having editorials about the foreshore or race relations, find stories about Don Brash and the foreshore or race relations.
This is almost the same as the previous query, but basically uses the newspaper editor as a relevance filter. It feels contrived; basically these two examples fail the "some coupling" requirement.

2S From movies in the detective story genre set in San Francisco, select scenes where Nicole Kidman speaks.
This satisfies the "sizeable fragment" requirement.

EC From CDs that contain Irish music, select planxties.
This satisfies "low coupling", "sizeable fragment", and "some coupling".

2S From movies whose sound track was composed or arranged by John Williams, select producer and director.
This shows that a meaningful query need not satisfy "sizeable fragment", but it is not an IR query, let alone an INEX one.

2S From books about anatomy, select sections about the articulation of the jaw.
This is a real query I had while I was writing the paper. The answers I found satisfied "low coupling" and "sizeable fragment".

2S From books about Bioinformatics published after 1994, select portions about Dynamic Time Warps.
Publication date is a property of the books as wholes, not of sections. Dynamic Time Warps have many applications other than Bioinformatics. So this satisfies "some coupling" as well as "sizeable fragments".

EC From books by Terry Pratchett, select chapters that mention a "Soul Cake" day. This illustrates the query-relative nature of coupling. Chapters are coupled to their contexts, but if all you want to know is which day of the week Soul Cake day is on, that does not matter. This is a small example of an information extraction query, suggesting that we should look to information extraction problems and collections for models.

EC From all Koine Greek documents in a collection of ancient documents, select paragraphs containing the word "$\pi\alpha\iota\sigma$". This is a real question I'd like to ask. It is a typical word study where the question is "how is this word used". The language, period, genre, even author of the documents could be relevant to the scope of the study. The fragments are, from a general point of view, tightly coupled to their context, but for the purposes of this kind of query, that semantic coupling is not relevant. Because it is concerned with a specific word rather than the meaning of the word, it is not really an IR or INEX query, so "low coupling" remains a desideratum for the INEX Question.

2S From R packages that are about trees, select function descriptions that are about pruning trees.
There are over 1200 pages of function documentation for core R; the contributed packages add about as much more. The function descriptions are similar to UNIX manual pages, only bigger. This satisfies "some coupling" and "sizeable fragments".

EC From volumes of Otago examination papers dated 2000 or later, find questions in COSC papers that mention Pascal. I have a DTD for this, and have personally marked up many COSC exam papers. I do not, however, have complete volumes, otherwise this would be a real question.
This satisfies "strong semantics", "low coupling", "some coupling", and "sizeable fragments", in that questions are a paragraph to half a page in size.

## References


1. Proceedings of the First Workshop of the Initiative for the Evaluation of XML Retrieval (INEX); December 9–11, 2002; Schloss Dagstuhl;
`http://www.ecrim.org/publication/ws-proceedings/INEX2002.pdf`
2. INEX 2003 Workshop Proceedings (2003); December 15–17, 2003; Schloss Dagstuhl;
`http://citeseer.ist.pesu.edu/649846.html`
3. INEX 2004 Workshop Proceedings (2004), this volume.
4. Narrowed Extended XPath I; Trotman, A., & Sigbjörnsson, B., (2004); in [3].
5. NEXI, Now and Next; Trotman, A., & Sigbjörnsson, B., (2004); in [3].