Microsoft Excel caBIG Smart Client Joining the Fight Against Cancer –

Katarzyna J Macura M.D. Ph.D.

Problem Statement:

The **ca**ncer **B**iomedical **I**nformatics **G**rid (***caBIG)*** https://cabig.nci.nih.gov/ is a voluntary, virtual informatics infrastructure that connects data, research tools, scientists, and organizations. Its goal is to speed the delivery of innovative approaches for the prevention and treatment of cancer.

caBIG was launched in February 2004 under the leadership of the National Cancer Institute's Center for Bioinformatics. Nearly 500 people from 50 NCI-designated Cancer Centers and 80 volunteer organizations and other interested groups are working on 70 ***caBIG*** related projects in a 20M dollar three-year pilot project.

caBIG is called the *World Wide Web of cancer research* by National Cancer Institute Andrew von Eschenbach, M.D. and it is envisaged as the solution to the *"Informatics tower of Babel"* that is stifling the cancer research community. An overwhelming volume of data is available from a multitude of sources in a variety of proprietary and incompatible data representation formats. Integration is seen as critical to achieving the promise of personalized medicine. By combining shared vocabulary, data elements, and data models **caBIG** is expected to become the *lingua franca: a* common-standard informatics platform used for all cancer research supported by NIH grants. This is a requirement resulting from the NIH's new policy on data sharing http://grants.nih.gov/grants/policy/data_sharing/.

It is very likely that ***caBIG*** will become a significant -- perhaps the authoritative -- source of variety of knowledge related to cancer. In Dr. von Eschenbach's own words, "today nothing is more critical, not even discovery of new gene or new medicine than productive use of the data tsunami generated by cancer research."

caBIG is based on four pillars: 1) *Open Source* -- all computer program source code is available at no cost to everyone, 2*) Open Development* -- **caBIG** 's goals, priorities, and activities are steered by public groups, five Workspaces and three Strategic Level Working Groups, that any interested party can join, 3) *Open Access* -- any person or group can connect to **caBIG** resources as loosely or as strongly they desire by conforming to **caBIG** technological standards because data has value beyond original purpose for collection, scientific methods demand verification by peers, and the obligation to share publicly funded

data products and 4) *Federation* -- local control of deployments. **Currently and mistakenly Microsoft products are consider incompatible to caBIG's pillars.**

Microsoft Office (MS) Excel's functionality, especially in terms of statistical analysis and visualization, allows it to dominate the spreadsheet market. These same features make MS Excel attractive to scientists in the Biomedical fields. For the last decade, the spreadsheet -- specifically MS Excel -- has been the primary manner by which biologists and biomedical scientists analyze cancer data.

Scientists traditionally analyzed a limited set of data collected in their labs or made available by a selected few collaborators. Thanks to the **caBIG** environment, scientists will have access to orders of magnitude more related data from researchers around the world. How will they analyze it to make meaningful deductions? **I have the controversial opinion that caBIG's current policy of developing vertical open-source tool packages from scratch with out liaising with established horizontal proprietary software is misguided**.

I strongly believe that extensions to MS Excel that allow users to access caBIG data-services could be efficiently developed with MS Visual Studios Tools for Office 2005 and will be widely used because they will leverage the scientists' intimate familiarity with MS Excel.

caBIG data-services are web applications that service XML formatted requests. The proposed ***MS Excel caBIG extensions*** will be a set of Windows Form Contextual GUIs for constructing appropriate XML queries and reformatting the serialized XML responses as cells in the Excel Worksheet.

This project would be very difficult to implement without the novel technology Microsoft is developing with Visual Studio 2005 and the .NET framework.

Visual Studio programming, debugging, and profiling tools -- especially the Excel integrated design-time experience -- facilitate very high developer productivity permitting this project to be developed at such low cost.

The .NET framework provides robust security, easy, managed deployment and rich support for XML. Managed .NET code allows updates (in the form of DLLs) to be detected and downloaded from servers but also allows the user to exercise strong control over how the code will be executed. The document-centric model means updates will not be a new version of the document, just a version of the DLLs.

Visual Studios Tools for Office permits us to stand on the shoulder of giants by developing on top of MS Office. It was said time and again by all already funded Project Directors who successfully implemented their projects that they only needed one programmer to do the job of 50 to 100 programmers (Drs. Patrick Hogan and Greg Quinn via webcast http://research.microsoft.com/workshops/FS2005/webcasts/12508/lecture.htm). This is in striking difference to 30 programmers who working for three years to develop Google Earth application.

Office's extendibility with the Document Actions Task Pane and Windows Form Contextual GUI allows the DLL to be accessible by scientists in an intuitive manner. MS Excel runs on a wide range of Microsoft Windows operating systems including Microsoft Mobile 5.0. This translates into **caBIG** access on a wide range of computing devices including PDAs and mobile phones. This "data intimacy" allows on-the-spot hypothesis testing no matter where is the spot; thus replacing the proverbial "restaurant napkin sketches".

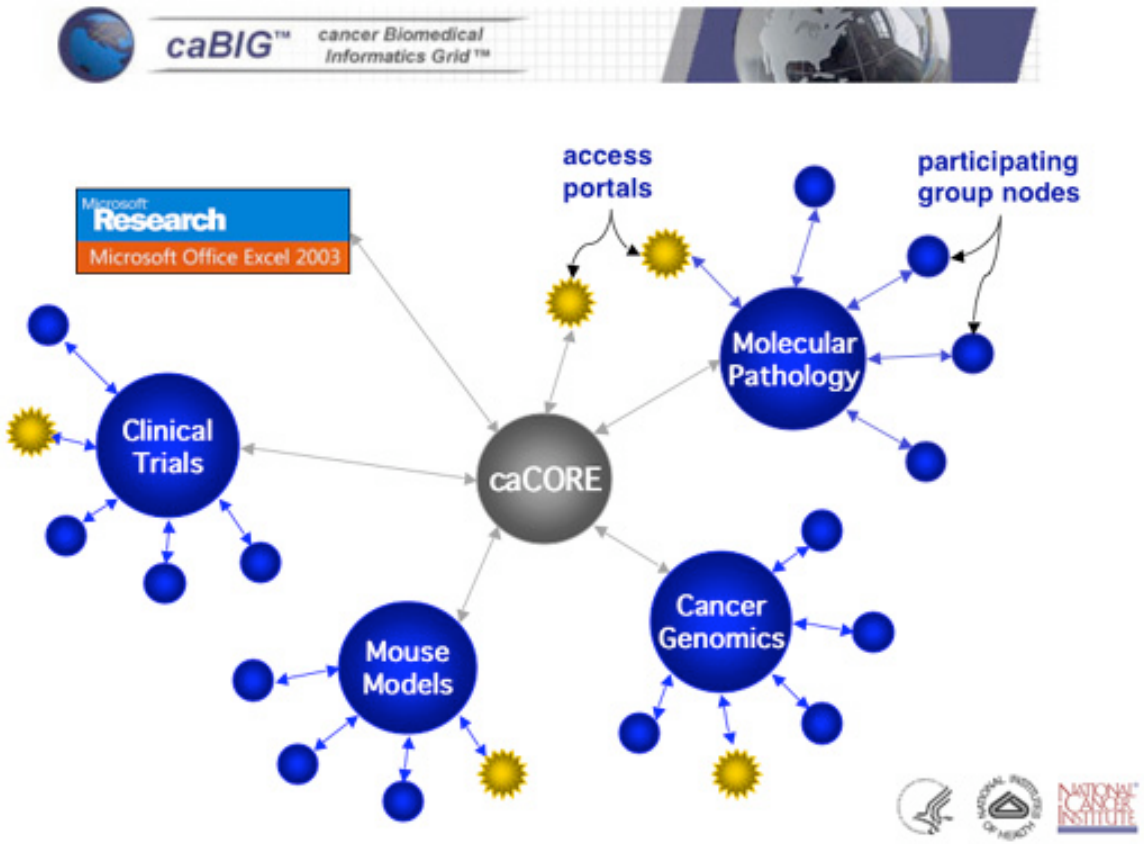

Figure 1: Proposed access to caBIG data-services in MS Excel

Schedule:

October 2005: Attend Microsoft Research eScience Workshop 2005 **http://research.microsoft.com/workshops/escience2005/,** discuss this proposal and learn from other researchers. Search for the developer starts.

Nov 2005: Notification of awards from Microsoft and MS Visual Studio 2005 is released. If funding is secured, I will construct a place holder website advertising the goals of this project and that a developer position is available.

Jan 2006: A developer is selected and hardware is bought. The first month will be spent, depending on the developer's strengths, gaining familiarity with MS tools and caBIG. The project website will be expanded and CVS, bug reporting, and public mailing-lists will be setup.

April 2006: Beta version of MS Excel extensions is released and made available for download on the website. This version will be tested by Johns Hopkins collaborators who I will personally recruit and any other interested pioneer users. Confirmed bugs and suggested enhancements will be collected.

August 2006: First public-ready version is released. We will attempt to interest a wide range of users from the **caBIG** community by advertising this project on the official **caBIG** web-sites and conferences. Most developer time will be spent on fixing bugs and supporting the user base through forums and mailing-list emails rather than on adding new features,

Oct 2006: Second public-ready version will be released and active paid development will be over. I hope that the product will have a sufficiently established user and development community that other interested **caBIG** members will take over maintaining and extending functionality. I am willing to continue leadership and oversight role past October 2006 as long as these extensions are useful to the community.

Expected outcomes

We will produce extensions to MS Excel as Windows Form Contextual GUIs that will allow users to access **caBIG** data-sources. The GUIs will transparently construct appropriate **caBIG** XML queries and reformat the serialized **caBIG** XML responses as cells in the Excel Worksheet.

For this project to mature to its full potential, it needs a strong user and developer community that

will support it. If this community succeeds, it will leverage the money Microsoft invested in this project by helping the paid developer find bugs, select worthwhile features, and expand functionality by contributing code. For users, this project's website will advertise the project, offer tutorials and be a distribution source for downloading the MS Excel extensions. Developer's will use our Tiki page, CVS, bug reporting, and public mailing-lists.

Microsoft can use this project as a Case Study demonstrating how Visual Studio 2005 for Office was used to expand Excel functionality in creative ways.

Use of Funds:

$45,000 USD Salary for Lead Programmer

The majority of the money will be used to fund a 9-month independent contractor position who will work full-time equivalent on this project for 36 weeks at $1,000 a week ($36,000) from about January 15th 2006 till October 15th 2006.

Hiring the developer as a contractor instead of a Johns Hopkins' employee avoids many Hopkins bureaucratic requirements and gives me much more flexibility over the person's employment.

The lead programmer must have degrees in computer science and mathematics (B.S equivalent or higher) in addition to knowledge of biology with experience in computational science research (e.g. biomedical engineering, bioinformatics). Fluency in grid programming and familiarity with Microsoft tools is required. Working knowledge of **caBIG** environment and development policies will be essential.

We will establish email/instant-messaging/tele-conferencing line of communication to unofficially discuss problems. Each week there will be 2-hour long progress meetings that will be minuted by the programmer.

I will use the remaining $9,000 (the difference between $45,000 and $36,000) as a source of performance bonuses for the programmer and to cover additional expenses such as communication costs, printing, poster design, and publishing.

$3,000 USD Hardware and Software

I would like to buy a laptop and desktop PC running Windows XP that will be used by the hired developer for writing and demonstrating the MS Excel extensions. The desktop PC will be used to host a

dedicated **caBIG** data-service that will be accessed by MS Excel running on the laptop. I would also like to purchase a mobile device running Windows Mobile 5.0 and MS Office to use for developing **caBIG** Excel extensions for thin-clients.

The developer will need a single license for MS Visual Studios 2005 and two Office 2003 Professional licenses. I hope that MS will be able to provide this software at no cost. If not, I will use this money to purchase the licenses.

I noticed that both Mr. Baker and Dr. Hogan had problems with live demonstrations during Microsoft Research Faculty Summit 2005. They blamed it on slow wireless connections. Unfortunately if similar incidents happened during our demonstrations to the medical and biomedical community, it would greatly blemish our project. Those communities have much less technical knowledge and therefore less understanding to such glitches. That is why we need two computers to create our own network.

$2,000 USD Travels

The remainder of the money will be used for to pay for the developer's travels to Microsoft and caBIG events to publicize and promote MS Excel **caBIG** extensions within the community of future users. I have travel funding for myself already secured.

I would like us to attend: (1) *caBIG Annual Meeting* [April 06], (2) a scientific conference such as *IEEE Computational Systems BioInfromatics* [August 06], and (3) Microsoft eScience Workshop 2006 [October 06].

Dissemination and Evaluation:

We will release software and its source code using a BSD open source license through the website (see section *Expected Outcomes*) All written materials such as tutorials will be released under the OPL (Open Content License). The project will be proactively shown to the biomedical community. We expect that Johns Hopkins's prestige and esteem in medical and scientific community will add credibility to the proposed work. We will evaluate our success based on how many people are using and contributing to this project and their feedback.

Other Support:

The Johns Hopkins University Department of Radiology and Radiological Sciences budgets 20% of my time to be spent conducting (unpaid) research projects. I will use part of this time to cover my salary for the duration of the proposed project. I will use departmental funds to pay for personal travel related to this project. Also, I have $25,000 unrestricted contract funds from **caBIG** portions of which I could use to supplement Microsoft funds.

Qualifications of Principal Investigator:

I am a practicing academic radiologist at the Department of Radiology and Radiological Sciences, Johns Hopkins University. I have Ph.D. degree in medical informatics (artificial intelligence). In a sense my background is very similar to both David's (Heckerman) http://research.microsoft.com/~heckerman/ and Eric's (Horowitz) http://research.microsoft.com/%7Ehorvitz/), both of whom I know. In addition to being certified by the American Board of Radiology (May 2000), I am also certified knowledge engineer (March 1993) by the International Association of Knowledge Engineers. Recently I was awarded one of only 15 contracts (450 researchers applied) by **caBIG** In Vivo Imaging Workspace. In short I have unique double training and experience that allows me to work on projects and with people from both the computer technology and medicine camps. You might say that I am bilingual. In fact I would say that I am bi-cultural. I, having been there and done some of it, can not only talk that talk but also walk a walk.

I truly believe that the **caBIG** is the Internet for Cancer Research and I would welcome the gift of being a part of the team that designs and implements this discovery engine of cancer processes. Adding creative and processing power of Microsoft tools will make this endeavor accessible to more people. I would be glad to contribute to the **caBIG** Excel project and I would be grateful for a chance to learn more from the experts in the field who are already involved in this monumental task.