Schlußbericht zum Projekt GlobalInfo – Gruppe 1, SFM2

CML-basierende interaktive dynamische Dokumente, neue Verfahren und Werkzeuge für elektronisches Publizieren

Teilprojekt: Neue Transferverfahren und Darstellungssysteme

Zuwendungsempfänger: Creon·Lab·Control AG, ehemals LabControl GmbH

Förderkennzeichen: 08SFL02 8

I. Kurze Darstellung

1. Aufgabenstellung

Um die Nutzbarkeit von Primärdaten in wissenschaftlichen Publikationen zu erreichen, müssen diese für den Anbieter wissenschaftlicher Publikationen erst einmal verfügbar gemacht werden. In einem zweiten Schritt müssen die Daten durch geeignete Werkzeuge in die Publikationen eingebunden werden können. Abschließend benötigt der Nutzer der Publikationen entsprechende Darstellungstools (Viewer) für die interaktive Nutzung der Primärdaten.

Ziele dieses Forschungsprojekts waren deshalb:

- Die Weiterentwicklung von Standardformaten für den Austausch und die Übermittlung von Primärdaten (insbesondere JCAMP-DX) für eine möglichst breite Palette von Datenquellen,
- Die Verbesserung der Kodierung von Primärdaten vom Autor zu den Anbietern wissenschaftlicher Publikationen,
- Die Erzeugung CML-basierter dynamischer und mehrschichtiger Dokumente und
- Die Entwicklung interaktiver Darstellungsprogramme für Web-Browser zur Nutzung der Daten.

Ein durchgängiges Konzept für die Entwicklung eines Autoren-Tools für das Aufzeichnen, Erfassen, Speichern, grafische Darstellen, Bearbeiten und Versenden (z.B. via Internet) von Spektren und Strukturen (einschließlich zugehöriger chemischer und technischer Informationen) war zur Erreichung dieses Zieles notwendig. Erste Ansätze fanden sich hierzu im Tool TranSpec der Arbeitsgruppe Moll. Dieses Konzept sollte konsequent so weiterentwickelt werden, daß die Gestaltung des Tools als *offenes System*, auf allgemeinen Prinzipien beruhend, die Bearbeitung auch nicht-spektroskopischer Daten ermöglicht.

2. Voraussetzungen

Die Creon·Lab·Control AG (ehemals LabControl GmbH) beschäftigt sich seit 1991 mit der Beratung, Standardisierung und Publikation von wissenschaftlichen Primärdaten, insbesondere aller Arten von Spektren

und zugehöriger Informationen auf elektronischem Wege. Creon·Lab·Control ist in diesem Zusammenhang weltweit als Dienstleister für pharmazeutisch-chemische Unternehmen, wissenschaftliche Gerätehersteller und wissenschaftliche Verlage tätig.

Durch enge Zusammenarbeit mit den Meßgeräteherstellern (u.a. Bruker, Shimadzu, Perkin-Elmer, Varian) wurde in der Vergangenheit die Generierung von Primärdaten standardisiert und somit für die elektronische Publikation vorbereitet. In Zusammenarbeit mit Verlagen (u.a. Wiley-VCH und Hanser) wurde anschließend die Publikation der Referenzdatensammlungen z.B. von Hummer (IR-Atlanten) in herkömmlicher Buchform und elektronischer Form (u.a. für SPECINFO) vorangetrieben.

Creon·Lab·Control befaßt sich außerdem intensiv mit der Weiterentwicklung von Standard-Datenformaten wie JCAMP-DX und JCAMP-CS in den entsprechenden internationalen Arbeitsgruppen (IUPAC Working Party on Spectroscopic Data Standards). Für JCAMP-DX enstand u.a. ein frei erhältlicher Spektren-Viewer „Spectacle ViewPoint" für zwei- und dreidimensionale Spektren. Spectacle ViewPoint wird als interaktiver JCAMP-DX-Spektren-Viewer von Webbrowsern in elektronischen Publikationen im Internet eingesetzt (in Zusammenarbeit mit Wiley-VCH/Chemical Concepts).

Creon·Lab·Control verfügt auch über selbstentwickelte Spektroskopiedatenstationen zur Erfassung, Visualisierung und Bearbeitung von Primärdaten: SPECTACLE-Produktfamilie. Schwerpunkte liegen hier auch auf der Standardisierung von Primärdaten durch Konversion aus proprietären Herstellerformaten.

Neben dem JCAMP-DX-Spektren-Viewer existiert der Prototyp eines intelligenten Web-Browsers mit integrierter ein- und mehrdimensionaler Spektren-/Struktur-Darstellung: SPECTACLE Chembrowser.

Weitere Ergebnisse vorangegangener Arbeiten sind Prototypen interaktiver JAVA-Applets für die Darstellung von Spektren und Strukturen in HTML-Dokumenten.

3. Planung und Ablauf

Die folgende Aufstellung enthält die einzelnen Arbeitsschritte für das Teilprojekt:

1. Die Weiterentwicklung der bereits genutzten Austauschformate für TranSpec, insbesondere das von fast allen Geräteherstellern genutzte JCAMP-DX, für andere bzw. verwandte Spektroskopie- und Datenarten (z.B. Fluoreszenz, CD, Kinetiken) wird evaluiert und Vorschläge für neue JCAMP-DX-Typen im Zusammenhang mit der reibungslosen Umsetzung in die Chemical Markup Language (CML) erarbeitet. Das JCAMP-DX-Format ist seit 10 Jahren eingeführt, es wird von den Meßgeräteherstellern unterstützt und deshalb sind besonders viele Daten in diesem Format verfügbar. Creon·Lab·Control ist Mitglied in der IUPAC Working Party on Spectroscopic Data Standards und kann Einfluß auf die Weiterentwicklung dieses Formats und die Umsetzung in CML nehmen. Das JCAMP-DX-Format bleibt auch nach der Umsetzung in CML für längere Zeit die Schnittstelle zum Mehrwert. Nur der bidirektionale Weg (Information kann in diesem Format leicht an den Nutzer zurückgeliefert werden) erschließt dem Wissenschaftler die Möglichkeit, elektronisch publizierte Information direkt in seine tägliche Arbeit einfließen zu lassen, z.B. die Spektren aus elektronischen Publikationen in seinen vor Ort eingesetzten Expertensystemen zu nutzen. Die bei der Umsetzung von JCAMP-DX in CML

für Archivierung und elektronische Publikation entstehenden Probleme werden analysiert und behoben.

2. Weiterentwicklungen der bereits genutzten Austauschformate für TranSpec, insbesondere JCAMP-DX, im Hinblick auf die Spektrum-Struktur-Korrelation (z.B. Assignment von Signalen zu Strukturfragmenten, speziell für die optische Spektroskopie und Massenspektrometrie) werden erarbeitet und definiert. Diese Erweiterung ist notwendig, da die bestehende Definition nicht computerlesbar ist. Die Computerlesbarkeit ist aber eine wesentliche Voraussetzung für den Transport dieser Information. Erst computerlesbare Definitionen lassen sich auch sinnvoll in interaktiven Viewern nutzen.

3. Die Anbindung von TranSpec an einen SPECINFO-Server (oder äquivalenten Server) zur automatischen Übermittlung der Publikationsprimärdaten vom Autor zum Verlag via Internet unter Verwendung von JAVA soll realisiert werden. Idealerweise ist ein Publikationstool nicht von der Plattform (dem Betriebssystem) des Computers abhängig. Ein JAVA-basiertes Werkzeug ermöglicht die Erstellung einer Publikation auf nahezu allen Betriebssystemen mit einer uniformen Software. Dies fördert die Verfügbarkeit und Akzeptanz.

4. Es sollen JAVA-Applets zur interaktiven Darstellung von Spektren und Spektrum-Strukturbeziehungen, unter Nutzung der unter Punkt 1 und 2 erstellten Erweiterungen, erstellt werden. Bisherige Applets verfügen noch nicht über eine interaktive Kopplung der stukturellen und spektroskopischen Information. Dies gilt insbesondere für TeleSpec und die von ACD Labs angebotenen Applets, die im Wesentlichen nur Strukturen oder Spektren darstellen. Die angebotenen Applets haben darüberhinaus Mängel in der Rücklieferung von Daten, z.B. TeleSpek liefert syntaktisch fehlerhafte JCAMP-DX-Dateien als Ergebnis einer Spektrenvorhersage.

5. Die unter 4. Erstellten Applets sollen in der Praxis in interaktiven Publikationen im Internet getestet werden. Hierzu gehört die Erstellung von Testpublikationen mit Anbindung an ein Host-System und Untersuchung dieser unter verschiedenen Aspekten, z.B. zur Lesbarkeit, Akzeptanz, Nutzung und zum Wissenstransfer.

4. Wissenschaftlicher und technischer Kenntnisstand

Suchbare Volltextversionen (einschließlich Grafiken) sind längst Standard der internationalen naturwissenschaftlichen Großverlage/Gesellschaften. Auf dem Gebiet des interaktiven Electronic Publishing existieren eine Reihe vielversprechender Ansätze, die es dem Autor aber bisher nicht ermöglichten, auf einfachem Weg seine experimentellen Primärdaten in die Publikationen einfließen zu lassen. So besteht eine große Diskrepanz zwischen publizierten und in Datenbanksystemen wie SpecInfo archivierten Spektren.

4.1 Internet

Bisher laufen die Bemühungen um Interaktion mit Daten im Internet. In den letzten Jahren wurden von der Arbeitsgruppe um Rzepa und Whitaker *Chemical Mime Types* vorgeschlagen und von der IETF als Standard anerkannt, die eine Einbindung chemischer Daten wie Molekülstrukturen, Reaktionsgleichungen und Spektren in HTML-Dokumenten ermöglichen. Zusammen mit anderen Datentypen, wie z.B. VRML, Shockwave, Toolbook und JAVA-Applets, sind damit interaktive Elemente innerhalb elektronischer

Publikationen möglich geworden. Um diese Informationen entsprechend visualisieren und bearbeiten zu können, wurden von den Forschern und kommerziellen Firmen kostenlose Viewer, Browser-Plugins und Java-Applets zur Verfügung gestellt. Teilweise sind diese Programme aber nicht für alle Plattformen erhältlich oder sie ermöglichen nicht alle erwünschten Analysemöglichkeiten und Interaktionen, so daß dringend Neu- und Weiterentwicklungen erforderlich sind.

4.2 Insellösungen

Nahezu jedes Hightech-Gerät und alle modernen Spektrometer sind mit hochinteraktiver gerätespezifischer Software ausgestattet. Sie dienen der Primärdatengewinnung und interaktiven Gewinnung publikationsfähiger Grafiken, Bilder, Tabellen, die dann allerdings nur noch einen geringen Teil der Informationen enthalten. Leider wählt jeder Gerätehersteller ein eigenes Datenformat und es besteht meist keine Querkompatibilität innerhalb desselben Arbeitsgebietes und schon gar nicht zu anderen Arbeitsgebieten. Dies schafft lokale Erfahrungen mit interaktiver Datenbearbeitung (Bilderzeugung), erschwert aber z.B. die Ausweitung von Datenbanken. Auf dem Gebiet der Spektroskopie setzte hier TranSpec (Moll) zu einer übergreifenden Lösung an.

Bildverarbeitungs-Software (wie die Software OPTIMAS) kann nur einen geringen Teil der Interaktionserfordernisse abdecken, wie das für viele Anwendungen in der Mikroskopie ausreicht. Die vielgestaltige Interaktion mit 3D-Objekten wie etwa komplexen Oberflächen (z.B. AFM, Rasterelektronenmikroskopie usw.) ist bisher nicht zwischen verschiedenen Geräten übertragbar.

4.3 Datenbanken

Aus publizierten Daten werden in den Naturwissenschaften mit großem Aufwand Datenbanken mit Spektren und Molekülinformationen erstellt. Obwohl JCAMP sich bei den Geräteherstellern etabliert hat und nahezu jede Meßsoftware das JCAMP-Format erzeugen kann, enthalten die Spektrendatenbanken leider nur einen kleinen Teil der ursprünglichen Information, da die Primärdaten in der Publikationskette verloren gehen. Ein Ansatz, dieser Entwicklung entgegenzuwirken, ist das in der Beta-Phase befindliche Autorentool TranSpec. Die weltweite Akzeptanz des Austauschformats JCAMP-DX ergibt sich aus der Breite der Anwendungen. So enthält die derzeit größte Sammlung kommerzieller Daten verschiedener Hersteller schon jetzt mehr als 700.000 Spektren im JCAMP-DX Format. Eine Umsetzung von JCAMP-DX Daten in CML mit Hilfe von TranSpec ist in der Entwicklung.

Die bereits professionell gelösten grafischen Suchmöglichkeiten in Chemie-Datenbanken betreffen nicht den Gegenstand dieser Sonderfördermaßnahme.

4.4 Publikationen

Angesichts der bisherigen Zurückhaltung von Verlagen (wirtschaftliches Risiko) und Förderprogrammen startete die Entwicklung im Internet. Der von Murray-Rust entwickelte, auf XML aufbauende Formatierungsstandard CML realisiert Zuordnungen von Informationen innerhalb eines Dokuments und schafft damit die Ausgangsbasis für die Print- und Onlineversionen sowie der Einbindung in Datenbanken. In CML ist die Einbindung von Metadaten, chemischen Formeln, Reaktionsgleichungen, 3D-Strukturen, Kristallstrukturen, Spektren und quantenchemischen Ergebnissen möglich. Zusammen mit Datentypen wie Shockwave, Toolbook, VRML und Java-Applets werden die interaktiven Elemente eingebracht. Zur Visualisierung und

Bearbeitung dieser Date gibt es kostenlose Browser-Plugins, Java-Applets und Viewer. CML ist um weitere Sprachmodule erweiterbar (MathML, TechML, ...), was für den interdisziplinären Ansatz wichtig ist. CML besitzt ein eigenes MIME-Format, und ein CML-Browser auf Java-Basis (JUMBO) ist in der Entwicklung. XML soll in zukünftigen Browsergenerationen von Netscape und MSIE unterstützt werden. CML als DTD von XML ist dadurch ein ideales, gut strukturiertes Format, um vielfach verlinkte heterogene Informationen zusammenzufassen. Hier kann aufgebaut werden. Der Wiley-Verlag (Schneider, Fröhlich) beginnt dies umzusetzen und erstellte Musterpublikationen mit externen Viewern. Rein elektronische Zeitschriften betonten bisher den Mehrwert der Primärdaten zu wenig und fanden nur geringe Akzeptanz, weil die Ausbildungsfragen (Autoren, Gutachter, Nutzer) zu wenig beachtet wurden.

5. Zusammenarbeit mit anderen Stellen

In enger Kooperation mit dem Projekt "Interaktive Publikation und Nutzung von 3D-Meßdaten" (Arbeitsgruppe Kaupp), welches sich mit der Einführung der Dateninteraktion beim elektronischen Publizieren mittels dynamischer Dokumente beschäftigt, wurde ein VRML-ASCII-Konverter entwickelt, um die bei Messungen mit einem AFM anfallenden ASCII-Daten in ein mittels VRML-Viewern und Plugins darstellbares und damit über WWW publizierbares Format (und wieder zurück) zu bringen.

Im Verlauf der theoretischen Arbeiten an diesem Projekt kamen verschiedene Fragen zum CML-Format auf, da CML nur für Strukturdaten eindeutig definiert ist. Die restlichen Datenarten sind nur recht „weich" definiert. Eine Zusammenarbeit mit den Erfindern von CML, Profs. P. Murray-Rust und H. Rzepa, kam jedoch leider nicht zustande, da diese Fragen zu CML prinzipiell nur auf Basis eines Beratungsvertrags beantworten. Dieses Vorgehen wird von der IUPAC mißbilligt.

CML ist bisher noch kein IUPAC-Standard und nutzt auch den Mehrwert, den XML bietet, nicht aus. Aufgrund des noch unsicheren Status von CML bei der IUPAC und der erwähnten unsicheren Definitionen für viele Datentypen wurde auf die Implementierung von CML als internes Datenformat vorerst verzichtet, das System aber gleichzeitig so flexibel angelegt, daß dies mit geringem Aufwand nachgerüstet werden kann.

Auf einem Treffen der IUPAC Working Party on Spectroscopic Data Standards im Juni 2001 wurden unsere Arbeiten zu einem XJCAMP-Format vorgestellt. Andere Stellen stellten Erweiterungen von XML vor, z.B. SpectraML (NIST).