文章编号: 1000-6281(2011)03-0215-07

扫描探针显微镜中数据格式的分析及转换程序的设计

黄龙飞,丁喜冬*

(光电材料与技术国家重点实验室,中山大学物理科学与工程技术学院,广东 广州 510275)

摘 要: 随着扫描探针显微镜技术发展和广泛应用,扫描探针显微镜的技术标准化问题在国内外受到越来越多的关注。然而在数据格式方面,由于所用硬件平台的不同等原因,各制造商所生产的扫描探针显微镜大多使用专用的数据格式。这些数据格式又往往互不兼容,给数据后续的分析处理和交流共享带来不便。本文首先对扫描探针显微镜中数据格式的发展及应用较多的格式进行了分析,然后从扫描探针显微镜现有数据的类型、内容和特点出发,研究了规范扫描探针显微镜数据格式的基本要求和思路; 经过比较和筛选提出了一种基于 HDF5 标准(hierarchical data format, HDF)的标准格式方案; 并用C++语言实现了对该格式的操作及其与几种专用数据格式之间的相互转换,完成了格式方案和代码资源的开源共享。

关键词: 扫描探针显微镜; 数据格式; 标准化; 格式转换; 开源

中图分类号: TH73; TP317; TG115.21+5.7 文献标识码: A

近年来,伴随着扫描探针显微镜(SPM)软硬件平台的不断发展和广泛应用,SPM 技术的标准化问题在国内外受到越来越多的关注[1~3]。国际标准化组织(ISO)已经于 2004 年将 SPM 标准化列入其工作框架之内并建立了相应的技术委员会和分委员会(TC201/SC3, data management and treatment)[4]。

在 SPM 的数据格式方面,由于所用硬件平台的不同等原因,各制造商所生产的扫描探针显微镜大多使用专用的数据格式。目前存在几十种不同的 SPM 数据文件保存格式,大部分都是非开放的,只能在某种特定的软件平台上存取及进行相关的处理。这些格式一般互不兼容,也难于转换,为数据后续的分析处理和交流共享带来不便。为此,ISO 建立了相应的研究小组(WG-1, information formats)深入研究 SPM 中的数据管理问题[4]; 其近期目标是发展一种公共的数据格式转换标准和相应软件使不同仪器采集的数据能够互相交换,提出 SPM 数据格式标准化的详细需求。根据 ISO 目前公布的基本思路,最终将建立集成的 SPM 数据库完成数据的标准化[5]。

本文研究 SPM 的数据格式及其标准化问题,提出一种基于二进制的数据存储方案,旨在建立一种能将 SPM 仪器所产生的数据及其处理结果进行完整存储的统一文件格式标准,实现存储格式的标准化。通过对 SPM 数据格式的广泛调查及目前发展趋势的研究,本文提出了一种基于 HDF5 的 SPM 统一数据格式规范方案。该方案能完全兼容现有流行的各种格式标准,并完全无损地将它们所包含的数据包容进来,实现了标准的前向兼容。本文所提出的方案能方便地实现现有格式向新格式的过渡,同时具有良好的开放性、可扩展性和存储性能,能够满足 SPM 领域未来对数据的存储、处理和共享的要求,可为 SPM 数据格式标准化的实现提供参考。

1 SPM 数据格式分析

1.1 SPM 数据文件及其组成

SPM 数据文件包含了 SPM 仪器对样品进行测量、扫描、加工等数据及后处理结果等相关数据,并以纯文件形式存储于普通文件系统中。一般地,SPM 数据文件所承载的主要信息包括两类: 实际数据及元数据。其中,实际数据一般是以多维数组形式存储的若干次彼此独立或相关的扫描、测量的数据。元数据主要是实际数据的相关注释或辅助说明,如数据的基本结构类型、数组的维结构、附加说明。除主要信息外,文件也可能会附带一定结构的文件头部。以上信息构成了 SPM 数据文件,如图 1

收稿日期: 2011-02-05; 修订日期: 2011-03-10

基金项目: 广东省教育部产学研项目(No. 2010B090400123).

作者简介: 黄龙飞(1982 -),男,硕士.

* 通讯作者: 丁喜冬(1968 -),男,博士,高级工程师、硕士研究生导师. E-mail: dingxd@mail.sysu.edu.cn

所示。

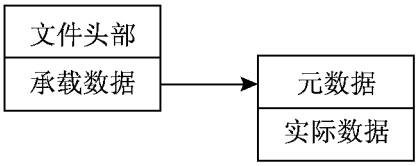

图 1 SPM 数据文件一般组成结构。

Fig. 1 General structure of SPM data file.

1.2 几种典型的格式

1.2.1 Veeco-NanoScope [6]

图 2 给出 Veeco 公司的 NanoScope 平台软件所生成的数据文件格式的典型结构。其中,元数据采用以行为单位的 ASCII 字符组织形成描述,图像等大量的数据附在文件最后并以二进制形式存储。数据文件后缀名通常可用正则表达式“[0~9]{3}”表示,从“000”开始自动递增。

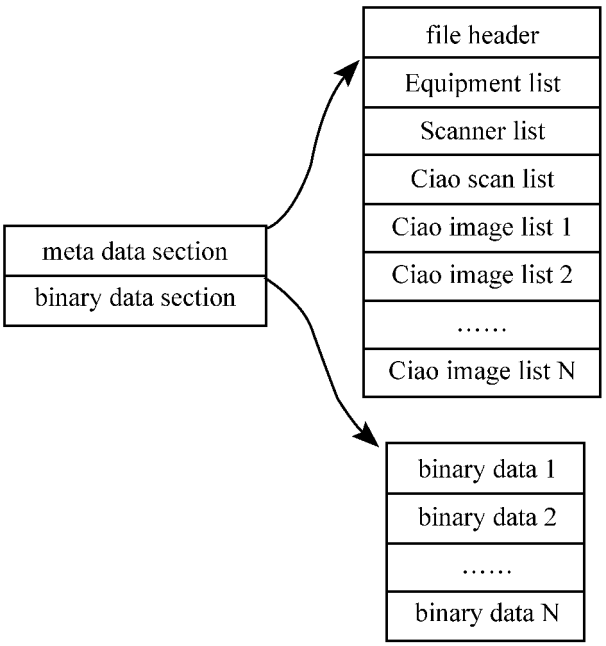

图 2 Veeco 格式文件构成 [6]。

Fig. 2 Structure of Veeco file [6].

1.2.2 RHK-sm3 [7]

RHK 公司 XPMPro 平台所创建的格式,由上一代标准 SM2 发展而来,文件后缀为“sm3”。其核心思想是将数据分页存储,每个文件由若干个数据页面组成,具有较好的层次性和良好的可扩展性,结构如图 3 所示。其中,数据的页类型分为图像、谱和注释 3 种。每个页面有其内部组成结构(包括头部和实际数据)。

1.2.3 BEING-csm [8]

中国科学院本原公司 CSPM 平台的数据记录格

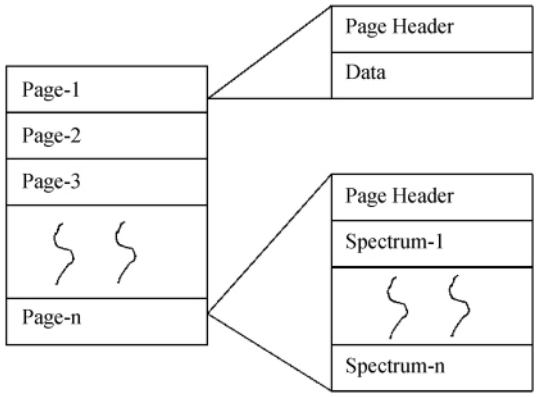

图 3 SM3 文件组成结构 [7]。

Fig. 3 Structure of SM3 file [7].

式,文件后缀为“csm”,其结构组成见图 4。

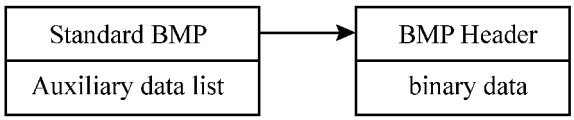

图 4 CSPM 文件格式组成结构 [8]。

Fig. 4 Structure of CSPM file [8].

其中,BMP 部分为标准的 32 位色深 BMP 文件格式组成的图像数据,实际上为 CSPM 平台采集的数字量。其后的辅助描述信息数据为 ASCII 文本按行存储。该格式以标准的 BMP 格式扩展而成,与绝大多数的图像处理软件兼容,易于处理。且基于 ASCII 的文件说明部分简单,容易提取出来,具有可读性。缺点是元数据说明部分信息量较少,并且较难实现较复杂的存储。

1.2.4 GXSM-NetCDF

GXSM [9] 采用 NetCDF(Network common data form) [10] 格式标准存储该平台的 SPM 数据。NetCDF 是大气研究大学协会(UCAR)开发的的一种跨平台的开放数据格式,文件后缀为“nc”。GXSM 通过定义合适的维度、变量来存储 SPM 所产生的多维数组类型的数据,并以属性存储元数据。该格式优点是采用原生的 NetCDF 底层存储机制,因此,它能充分利用 NetCDF 的优点,构建复杂的数组结构,而属性类型则基本能完成各类元数据的存储需要。缺点是原生的 NetCDF 底层存储能支持的数据类型较为有限,且不能自定义,每个文件只有一个可拓展维数的记录类型。

1.2.5 SPML

Thijs Bolhuis 等开发的 SPML [11],采用 XML 格式进行数据存储,对于较少量的数据尚可,特别是基于文本对于元数据的自明性非常有利,但对于大量、

复杂的数据存储则缺点较为明显。

1.2.6　SYSU-NSPM

中山大学 SPM 研究组基于自行开发的 NSPM 平台采用的文件记录格式,将同一次测量过程主要的二进制数据文件和相关的说明性数据以相同文件名和不同后缀存储在不同的文件。其中实际数据采用 Little endian 二进制存储,元数据以 ASCII 文本分行存储。优点是可以通过不断增加数据文件的类型来扩展数据记录的类型。但同时该格式存在文件分散凌乱、不便于对大量数据处理等问题。

1.3　HDF5 简介

1.3.1　HDF5 由来及发展历史[12]

HDF(hierarchical data format)是一个开源项目,于 1987 年由美国国家超级计算应用中心(national center for supercomputing applications,NCSA)的图像基金会特别组(graphics foundations task force,GFTF)创建。支持编程语言包括 C++、Fortran、Java 等主流编程语言,以类似于 GPL 的伊利诺大学协议(University of Illinois license)发布,任何人均可通过互联网无偿获得 HDF 相关资源并在遵循协议的前提下进行开发利用。HDF5 是 HDF 的最新版本,也是重要的里程碑版本。

1.3.2　HDF5 文件的组成结构

HDF5 是一种分层数据格式[13]。基于 HDF 的文件是分层结构化的。其基本实现依赖组(Group)及链接对象(link objects)。组作为有向图节点,内部记录了多个链接,每个链接指向一个被命名对象,被命名对象可能是组、数据集或数据类型,通过链接指向形成有向图,从而形成复杂的结构。在此,本文只讨论严格遵循层次性(即类似于文件系统)的情况,回路不作讨论。

HDF5 将操作系统文件系统的组织思想应用到了单个文件中的内部结构中。它采用了 UNIX 系统文件目录树的方法来组织文件中的数据。如图 5 所示,其中组(Group)相当于文件夹,数据集(Datasets)相当于文件。

HDF5 的内部结构具体说明如下:

(1)组(Group)

组相当于 UNIX 文件系统的目录。一个组包括一个或多个子对象,而每一个对象必须至少是一个组的成员(根组除外,根组不属于任何组)。组的成员关系实际上是通过链接对象来实现的。

(2)数据集(Datasets)

数据集是由数据和元数据组成。HDF5 中的数据支持多种数据类型甚至复杂的组合数据类型。从应用的角度来看,数据通常由某种特定的数据类型的数据元素组成,以单维或多维数组的形式存储在数据空间(Dataspace),即实际的数据存储区内。此外,属性解释说明数据,属于元数据的一部分。

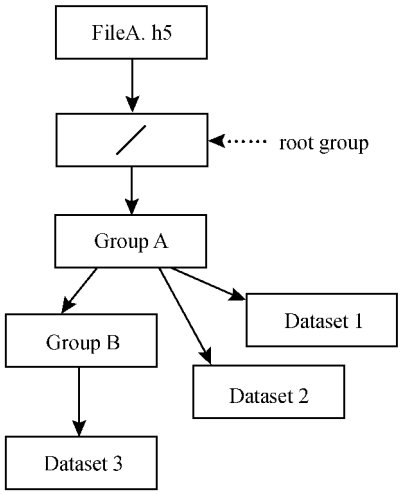

图 5　严格层次化组结构组成的 HDF 文件[13]。

Fig. 5　An HDF5 file with a strict hierarchical group structure[13].

(3)属性

属性用来说明数据集,实际上它本身是一种小数据集,但不能单独存在,隶属于所解释的对象,并直接存储于其头部。

1.3.3　HDF5 文件的特点

HDF5 充分借鉴了文件系统的组织方式,结构易于理解,体系开放,且其丰富的数据类型支持能满足各类科学数据的存储要求。同时,HDF5 从原理、代码、授权、文档、工具都具有充分的开放性,并且在各类流行语言上都有移植,易于实现具体的应用。因此,本文采用其作为标准的公共 SPM 数据文件格式的基础。

1.4　SPM 格式标准的选择

一个开放的公共 SPM 数据文件格式标准首先应该具有较好的兼容性,发展相对成熟,具有代表性。随着 SPM 向高速、大面积、高分辨率的趋势发展,意味着 SPM 数据有海量化的趋势。同时,未来有可能要求对数据进行实时高效的处理甚至并行处理,因此,存取性能和并行支持也是未来 SPM 数据文件的重要要求。HDF5 格式标准能够完全满足以上要求。以下将详细讨论一种基于 HDF5 的具有良好开放性的 SPM 数据文件格式标准 SPMDF(data file format,SPM)。

1.5 SPM 数据格式的转换

实际应用中经常需要将一种特定格式的 SPM 数据文件转换成另外一种格式，并保证转换过程不会导致数据的失真和丢失。由于目前存在的格式大部分为商业闭源软件所生成，其格式标准一般不公开。解析这类非开放格式标准的文件需要耗费代价，且由于软件版权保护原因，逆向工程受到限制；此外，格式标准制定者（一般为软件提供商）对格式版本方案可能进行频繁的升级，使得外界对私有格式文件解析项目的维护也带来困难。

另外，由于不同标准对诸如元数据定义、文件头规定、字符类型、文件最大尺寸、原生数据类型等都有较大区别，因此，对于任意指定的两种格式，转换过程一般不能保证 100% 的兼容，即数据完整性不能得到确保。转换数据格式的最终目的是为了正确解析或处理某一特定格式的数据文件。从根本上来说，只有开放、统一、具有广泛兼容性的公共标准才能消除以上问题。

2 SPM 数据格式和接口的规范

2.1 SPM 数据格式的规范

通过对目前若干种流行格式标准的分析，作者发现，“按测量或处理事件对 SPM 数据进行分区存储”的思想普遍应用。分区形式上可以是列表（对于 Veeco-NanoScope）、页（对于 RHK-sm3）、变量（对于 GXSM-NC）等。而在 HDF5 中，逻辑上很容易将分区映射为组，分区下的数组型数据则映射为数据集，元数据映射为属性。因此，可以通过这种逻辑映射储存复杂的 SPM 数据。

作者把每次典型的 SPM 使用过程（如测量或扫描）都看成一个事件（Event），每个事件对应着一个事件环境（Envirement），和若干数据（Data），即以事件为索引，环境和数据为核心，事件、环境、数据这三者包涵了当下及未来记录一个典型的 SPM 数据所需要的各个方面。综合 HDF5，作者建立了 SPMDF 的组成体系。一个典型的文件实例结构如图 6 所示。在这个文件中，根组下存在两个重要的子组，一个是事件（Event）组，另一个是环境（Envirement）组。事件可分成两类：一是数据源事件，包括测量、扫描等；二是数据处理事件，包括计算、变换等。事件在 SPMDF 中都以组数据处理事件依附于被处理事件，并作为其子组存在。SPMDF 以事件为脉络，以数据为核心，以属性为辅助，通过引用共享环境或数据。

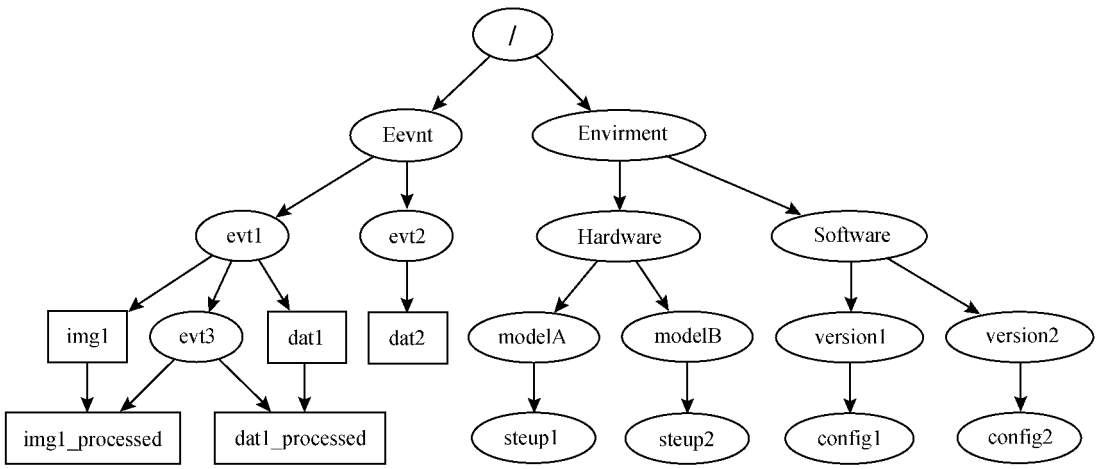

图 6 SPMDF 文件结构组成实例。

Fig. 6 An example of SPMDF file structure.

事件具有数据、相关环境及其它说明参数，如时间、持续长度等。每次事件在 HDF5 中作为一个组存在。组与组之间，由于是分立的事件，故彼此数据不发生直接联系，但可以共享相同的环境。在图 6 中，Event 组下有两个事件 evt1 和 evt2，它们可以共享相同的环境。事件环境对应着软件环境、硬件参数等相关参数，分别位于根组下面的环境组（Envirement）下。每个事件组中都有一个引用指向对应环境的组。这样，不同事件可以共享同一套环境参数，如不同测量过程中使用同种硬件可以共享同一套设置。

数据在 SPMDF 中是指实际的数据，可分为两大类：一是原始数据；二是处理后数据（包括图像、曲线等）。对于后者，其路径可以链接到其原始数据之下。数据作为事件的核心，以数据集的形式依附在事件组之下，不能脱离事件而存在。一个事件可

能有多个数据或无数据(如 SPM 处于工作中,数据采集尚未完成)。在图 6 中,dat1 和 dat2 分别是 evt1 和 evt2 的一个数据。图像 img1 由 dat1 处理而得,img1_processed 由 img1 处理而得。

针对 SPM 数据的特点,SPMDF 在 HDF5 的基础上作了简化。如: 限制 HDF5 中的对象只有一个父链接对象,即单亲链接; 规定链接必须为硬链接; 同时,组成 SPMDF 的组、数据集、属性须满足一定的命名规范,通过规范的命名空间和数据类型定义,从而可形成统一标准。

2.2　数据接口的规范

SPMDF 的接口基于 HDF5 的接口,是 HDF5 的包装层。这样,一方面可以充分利用 HDF5 底层的优点; 另一方面,又可以简化设计,屏蔽额外的特性,更加适合于 SPM 数据的储存。接口的体系如图 7 所示。部分典型的接口如表 1 所示。

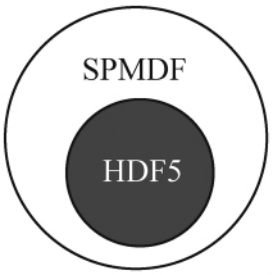

图 7　SPMDF 接口体系。

Fig. 7　Interfaces of SPMDF.

表 1　SPMDF 的部分接口

Table 1　Examples of SPMDF interface

主要关联对象	接口
事件	int CreateEvent(const string& farther,const string& event)
事件数据	int AddEventData (const string& event, hid _ t dataset)
数据属性	int SetDataAttr(hid_t dataset,hid_t attr)
环境	int CreateEnvirment (const string& farther, const string& envirment)
环境子项	int SetInstrument (const string& envirment, const string& instrument)

3　SPM 数据转换程序的实现与应用

3.1　格式转换的目的和原理

基于 HDF5 的 SPMDF 具有以下有用的特性: 具有各种数据类型的广泛支持,同时支持自定义数据类型,解决了数据类型的兼容性问题; 类文件系统的组成结构设计具有较强的可拓展性,可任意加入新的数据元素; 支持非常大的文件尺寸,能兼容现有的大容量文件。因此,SPMDF 能够无损地接受现有的任意格式的 SPM 数据。

利用这一原理可以实现任一格式到 SPMDF 的无损转换。同时,也可以将转换后的 SPMDF 文件无损转换回原格式,见图 8。将文件从 X 格式转换到 SPMDF 格式,首先要解析 X 格式文件,完全读取其元数据及实际数据。然后建立对应的 SPMDF 文件。最后转换 X 格式的数据类型到 SPMDF 内定的数据类型,并转换其变量名并按规则存入相应的路径。进一步,作为一个重要应用,SPMDF 能够作为数据文件的中间文件,在格式转换过程起到数据桥梁的作用,从而实现任意不同格式不间的转换。图 9 表示一个典型的转换过程。

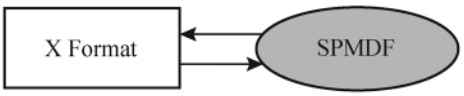

图 8　SPMDF 兼容性。

Fig. 8　Compatibility of SPMDF.

图 9　利用 SPMDF 进行格式转换。

Fig. 9　Format transform using SPMDF.

欲转换不同格式的文件,可先解析 A 格式文件,并将其无损转储到 SPMDF 格式,然后将自明的 SPMDF 文件转储为 B 格式。这个过程的信息完整性能否保持,依赖于 A、B 两种格式的兼容性。同理,也可实现 B 到 SPMDF 的无损转换和 B 到 A 的反向转换。

3.2　格式转换的特点

传统的格式转换一般只支持一对一的转换。考虑在一个传统的转换系统中,已有 N 种格式,为了实现格式间的互相转换,需要实现 2C2 N 个接口; 而采用上述以 SPMDF 这种公共格式为桥梁的转换,则只需实现 2N 个接口; 同时,当有新的格式加入时,传统的转换系统需要增加 2N 个接口实现,而上述转换则无须增加或更改之前的接口实现,只须实现 2 个转换接口即可。这样,大大增强了系统的可维护性。更重要的是,由于公共格式起着转换中枢的作用,必然要求它能够完全包容非公共格式的特性,有利于非公共格式的特性向公共格式进行迁移,从而逐步发展起一种开放、兼容的公共格式体系。当公共格式发展成熟时,让其完全代替非公共格式。

所以,这种体系设计,既兼容了非开放标准,又通过转换功能促进和引导了统一化的进程,符合当前 SPM 数据格式标准发展的要求。

3.3 实现思路和程序框架

以 C++下的实现为例,为实现上述功能,本文将上述转换的源码实现框架分为两个独立的部分。一是 SPMDF,二是非 SPMDF 标准的模块支持。其中,后者定义了一个重要的纯虚基类 SPMFileBase,其定义如图 10。

```
class SPMFileBase
{
    virtual int Convert2SPMDF(const string& filename,int mode)=0;
    virtual int LoadFromSPMDF(const string& filename,int mode)=0;
};
```

图 10 SPMFileBase 定义。

Fig. 10 Definition of SPMFileBase.

任何支持与 SPMDF 双向转换的非 SPMDF 文件格式均从这个基类派生,且必须实现 Convert2SPMDF 及 LoadFromSPMDF 接口。非 SPMDF 格式支持模块类关系体系设计实例如图 11 所示。图中 Veeco_NanoScopeFile,RHK_SM3File 分别为 Veeco-NanoScope、RHK-XPMpro 的格式文件类。通过实现它们的 Convert2SPMDF 与 LoadFromSPMDF 接口,即可实现与 SPMDF 的转换功能及彼此的相互转换。图 12 代码片断演示了 Veeco 和 RHK 两种格式之间的相互转换过程。

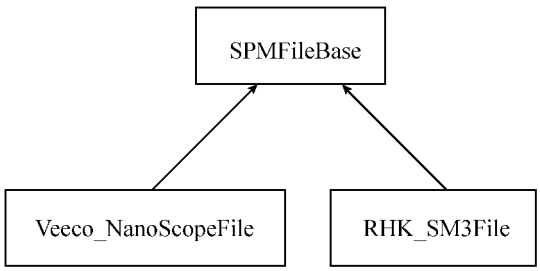

图 11 非 SPMDF 格式支持模块类关系体系设计。

Fig. 11 Non-SPMDF format moudule class view.

3.4 程序的开源与应用

本程序在 Windows 和 Linux 平台下得到实现和应用,目前支持的数据格式包括上述的几种格式。本程序采用 GPL 协议(general public license)开源发布,可以在 sourceforge 的本项目专属网站(http://sourceforge.net/projects/spmdf)任意下载使用。用户可选择合适的格式解析模块进行转换,或者自行加入新的模块以添加某种格式的支持。可以预见,随着格式支持模块的增多,转换功能将同步增加。同时,逐步丰富 SPMDF 的定义,扩充其数据包涵范围,使其逐渐成长为一个有代表性的开放性格式标准。

```
typedef Veeco_NanoScopeFile VeecoFile;
typedef RHK_SM3File    SM3File;

//read source file respectively
VeecoFile veeco_src("veeco.000","r");
SM3File rhk_src("rhk.sm3","r");

//create new file respectively
SM3File r_from_v("r_from_v.sm3","w");
VeecoFile v_from_r("v_from_r.000","w");

//convert from Veeco to RHK
veeco_src.Convert2SPMDF("tmp1.spmdf");
r_from_v.LoadFromSPMDF("tmp2.spmdf");

//convert from RHK to Veeco
rhk_src.Convert2SPMDF("tmp2.spmdf");
v_from_r.LoadFromSPMDF("tmp2.spmdf");
```

图 12 转换实现过程示例。

Fig. 12 A sample code for format transfor.

4 结论

标准化是 SPM 技术发展过程中的必然趋势,也是当前备受关注的问题之一。本文首次将基于 HDF5 的 SPM 文件格式标准应用于 SPM 领域。在国际、国内众多的 SPM 数据格式中,选择了几种有代表性的 SPM 数据格式进行了研究,分析了 SPM 数据的组成内容、结构和效率等特性。在此基础上,提出了一种基于 HDF5 的开放的格式方案即 SPMDF,实现了几种常见格式向 SPMDF 的转换。进一步地,以这种格式标准为桥梁,实现了不同格式之间进行转换,从而验证了 SPMDF 良好的双向兼容性和开放性。

下一步可继续探索更多现存 SPM 文件格式与 SPMDF 的融合,补充、丰富和完善命名约定规范,使 SPMDF 能在更大范围内得到推广并最终实现 SPM 数据格式的标准化。

参考文献:

[1] Fujita D, Itoh H, Ichimura S, et al. Global

standardization of scanning probe microscopy [J]. Nanotechnology, 2007, 18: 084002

[2] Fujita D, Onishi K, Xu M S. Standardization of nanomaterials characterization by scanning probe microscopy for societal acceptance [J]. Journal of Physics: Conference Series, 2009, 159: 012002.

[3] 王春梅, 井藤浩志, 孙洁林, 等. 研究开发适用于国际标准的 SPM 探针特性表征结构 [J]. 电子显微学报, 2007, 26: 576 - 581.

[4] Powell C J, Shimizu R. Development of standards for surface analysis by ISO technical committee 201 on surface chemical analysis [J]. Surface and Interface Analysis, 1997, 25: 860 - 868.

[5] ISO 14976. Surface chemical analysis-Data Transfer Format [S]. 2000.

[6] http://www.di.com/ [EB/OL]. 2009.

[7] XPMPro User Manual. http://www.rhk-tech.com [EB/OL]. 2009.

[8] http://www.spm.com.cn [EB/OL]. 2009.

[9] Horcasl I, Fernán-dez1 R, Gómez-Rodríguez J M, et al. WSXM: A software for scanning probe microscopy and a tool for nan-otechnology [J]. Rev Sci Instrum, 2007, 78: 013715.

[10] http://www.unidata.ucar.edu/software/netcdf/ [EB/OL]. 2009.

[11] Thijs Bolhuis, SPML White Paper. http://www.el.utwente.nl/smi/views/Research/Microscopy/Software/SPML/ [EB/OL]. 2010.

[12] http://www.hdfgroup.org/about/history.html [EB/OL]. 2010

[13] HDF5 User Guide. http://www.hdfgroup.org/HDF5/ [EB/OL]. 2010.

Standardization and transform of SPM data formats

HUANG Long-fei, DING Xi-dong*

(State Key Laboratory of Optoelectronic Materials and Technologies, School of Physics and Engineering, Sun Yat-sen University, Guangzhou Guangdong 510275, China)

Abstract: Standardization of SPM (scanning probe microscope) is drawing more and more attention nowadays. Most data formats in SPM are not compatible to each other, leading to the inconvenience of their subsequent analysis, processing and sharing. It is necessary to establish an open data format which has good compatibility and scalability. This paper analyzed the development of SPM data formats and some typical formats, and studied its standardization. Based on HDF5 (hierarchical data format, HDF), a solution for the standardization was given. A realization with C++ programming language was demonstrated. The solution and all related codes were published via internet, which were available and free to anybody.

Keywords: scanning probe microscope(SPM); data format; standardization; transform; open source

* Corresponding author