ICS 35.240
CCS A92

SF

中华人民共和国司法行政行业标准

SF/T 0158—2023
代替 SF/Z JD0403001—2014

软件相似性鉴定技术规范

Technical specification for examination of software similarity

2023－10－07 发布　　2023－12－01 实施

中华人民共和国司法部　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件代替SF/Z JD0403001—2014《软件相似性鉴定实施规范》，与SF/Z JD0403001—2014相比，除结构调整和编辑性改动外，主要技术变化如下：

a) 更改了“范围”的内容（见第 1 章，2014 年版的第 1 章）；
b) 增加了“规范性引用文件”（见第 2 章）；
c) 更改了“术语和定义”（见第 3 章，2014 年版的第 2 章）；
d) 更改了“仪器设备”（见第 4 章，2014 年版的第 3 章）；
e) 增加了“总体要求”（见第 5 章）；
f) 更改了“鉴定步骤”（见第 6 章，2014 年版的第 4 章）；
g) 更改了“鉴定记录”（见第 7 章，2014 年版的第 5 章）；
h) 更改了“鉴定意见”（见第 8 章，2014 年版的第 6 章）。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由司法鉴定科学研究院提出。

本文件由司法部信息中心归口。

本文件起草单位：司法鉴定科学研究院、公安部第三研究所、最高人民检察院检察技术信息研究中心、中国科学院软件研究所、国家工业信息安全发展研究中心、上海市人民检察院。

本文件主要起草人：郭弘、吴松洋、李佳、李岩、田野、丁丽萍、潘妍、高峰、卢启萌、杨恺、曾锦华、耿浦洋、李致君、毛晓、凌嵘、邹晓晨、方海峰、金波、张颖。

本文件及其所代替文件的历次版本发布情况为：

——2014 年首次发布为 SF/Z JD0403001—2014；
——本次为第一次修订。

软件相似性鉴定技术规范

1 范围

本文件规定了软件相似性鉴定的总体要求以及仪器设备、鉴定步骤、鉴定记录和鉴定意见的要求。

本文件适用于司法鉴定领域中对软件相似性的鉴定。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 11457 信息技术 软件工程术语

GB/T 29361 法庭科学 电子数据文件一致性检验规程

SF/T 0105 存储介质数据镜像技术规程

SF/T 0157 移动终端电子数据鉴定技术规范

3 术语和定义

GB/T 11457、GB/T 29361、SF/T 0157界定的以及下列术语和定义适用于本文件。

3.1

检材 questioned software/code for examination

电子数据检验鉴定中需检验的软件的程序、代码或开发测试文档等。

注：通常为涉嫌侵权的软件。

3.2

样本 known software/code for comparison

电子数据检验鉴定中用于同检材（3.1）进行比对检验的软件的程序、代码或开发测试文档。

注：通常为被侵权的软件。

3.3

数字水印 digital watermark

一种运用计算机算法嵌入载体文件的保护信息。

[来源：GB/T 36303—2018，3.1]

3.4

非原创文件 non-original file

不享有著作权的程序文件。

示例：公共程序库文件、第三方库文件和基于开源许可证的文件等。

4 仪器设备

4.1 硬件

鉴定所用仪器设备硬件宜包括但不限于：

a) 电子数据鉴定工作站；

b) 数码照相机；

c) 数码摄像机；

d) 存储介质只读设备；

e) 存储介质复制设备。

4.2 软件

鉴定所用仪器设备软件宜包括但不限于：

a) 完整性校验值计算软件；
b) 电子数据相似性比较软件；
c) 安装程序解包软件；
d) 计算机/移动终端系统仿真软件；
e) 网络数据包分析软件；
f) 反汇编分析软件；
g) 应用程序功能分析软件；
h) 数据库分析软件。

5 总体要求

5.1 排除非原创：如检材和样本包含非原创文件/内容，软件的相似性比对宜包含排除非原创文件/内容后的比对检验。

5.2 对等比对：检材和样本应在对等形态下进行比对检验，如源程序与源程序进行比对、目标程序与目标程序进行比对、文档与文档进行比对、数据库与数据库进行比对。如果检材与样本形态不一致，则应通过编译或者反汇编等手段转换一致后进行比对。

5.3 考虑差异：比对过程宜充分考虑名称（如文件名、变量名和函数名等）、语法（如空行、空格和字符大小写等）、注释、顺序以及编译器等差异造成的影响。

5.4 操作可追溯：与鉴定有关的情况应及时、客观、全面地记录，对于不可再现情况录像记录，应确保鉴定过程和结果的可追溯性。

6 鉴定步骤

6.1 记录检材/样本的情况

根据检材/样本的载体类型，应选择以下适当方式进行记录。

a) 载体为实物：对检材/样本的实物载体进行唯一性编号后拍照或录像，并记录其类别、品牌型号、唯一性编号、性状等。
b) 载体位于网络：记录检材/样本的下载网址、远程访问的账号及口令等信息。必要时，通过网络数据包监测，分析检材/样本的下载来源。

6.2 提取固定检材/样本

6.2.1 根据检材/样本的载体类型，应选择以下适当方式进行数据提取。

a) 载体为具备复制条件的数字化设备：按照 SF/T 0105 的规定制作镜像并核对其完整性校验值，然后采用只读方式对该镜像进行检材/样本的数据提取。
b) 载体为不具备复制条件的电子设备（非移动终端）：使用录像或者录屏的方式记录检材/样本的数据提取过程。
c) 载体为移动终端：按照 SF/T 0157 的规定对检材/样本进行数据提取。
d) 载体为网络：使用录像或者录屏的方式记录检材/样本的数据提取过程,并从可信时间源获取并记录开始时间和结束时间。

6.2.2 对提取的检材/样本应计算完整性校验值。

注：对于使用录像或者录屏的方式记录检材/样本的数据提取过程的，录像或录屏包含检材/样本的完整性校验值的计算过程。

6.3 相似性比对分析

6.3.1 相似性比对

6.3.1.1 基本要求

审查检材和样本的类型，选择合适的鉴定项目进行比对，常见的比对项目应包括但不限于：

a) 源程序的比对；
b) 目标程序的比对；
c) 源程序和目标程序的比对；
d) 开发测试文档的比对；
e) 开发测试文档和源程序/目标程序的比对。

6.3.1.2 源程序的比对

应按照GB/T 29361的规定分别对检材和样本中的源程序进行比对检验。若所有对应文件的完整性校验值相同，则软件相同；若对应文件的完整性校验值不同或存在无法对应的文件，则宜在排除非原创的内容后，对检材和样本的源程序的目录结构、文件名、文件内容、变量、函数和宏定义等分别进行比对检验。

6.3.1.3 目标程序的比对

应按照GB/T 29361的规定分别对检材和样本中的目标程序进行比对检验。若检材与样本程序文件相同，则软件相同；若检材与样本程序文件存在不同或存在无法对应的文件，则根据检材与样本的实际情况，选择以下1项或多项进行。

a) 安装程序检验：对检材和样本的安装程序（或解包后）的目录结构、目录名和各组成文件的文件名、文件完整性校验值、文件内容、文件结构、文件属性和文件签名信息等进行比对检验。
b) 安装过程检验：在相同环境下，分别隔离运行检材和样本的安装程序，对安装过程的屏幕显示、软件信息、安装选项和安装步骤进行比对检验，必要时也可对安装过程中的网络访问情况和调用程序库等进行检验和比对。
c) 安装后的程序检验：对安装成功的检材和样本的程序进行以下比对检验：
    1) 安装后产生的目录结构及目录名；
    2) 安装后产生的文件的文件名、文件完整性校验值、文件内容、文件结构和文件属性等；
    3) 安装后注册表的变动；
    4) 安装过程中产生的临时文件；
    5) 安装后软件的配置过程和运行方式；
    6) 软件使用过程中的屏幕显示、功能、功能键和使用方法等；
    7) 软件卸载过程中的屏幕显示和功能键等；
    8) 卸载后软件的残留文件；
    9) 卸载后注册表的变动；
    10) 程序逆向分析：目标程序如具有防检测分析的保护，如加壳和加密等情况，可根据需要先去除保护，再对目标程序进行反汇编,对反汇编后的代码应按照 6.3.1.2 的规定进行比对检验。

6.3.1.4 源程序和目标程序的比对

将源程序编译成目标程序后,应按照6.3.1.3的规定进行比对检验，或将目标程序进行反汇编后，按照6.3.1.2的规定进行比对。若目标程序具有防检测分析的保护，如加壳和加密等情况，可根据需要去除保护，再对目标程序进行反汇编。对于源程序编译过程中,由于编译软件和编译环境等不同而导致的文件差异，应进行记录。

6.3.1.5 开发测试文档的比对

应对检材和样本中的开发文档、需求说明书、设计方案、操作手册、开发手册和测试手册等相关文档的文本内容、流程图和属性信息等进行比对检验。对于包含文字的非文本形式的文档（如扫描文档），可进行文字识别并验证后进行比对检验。对于包含图片等多媒体数据的文档，应对多媒体的元数据和内容等进行比对检验。

6.3.1.6 源程序/目标程序和开发测试文档的比对

应对检材和样本中源程序/目标程序的作者/开发者和代码等信息与检材和样本中开发文档、需求说明书、设计方案、操作手册、开发手册和测试手册等相关文档进行比对检验。

6.3.2 相似性分析

6.3.2.1 总体分析

软件相似性鉴定中应了解与鉴定有关的情况，并对检材和样本整体相似情况进行分析，包括文件数量、文件类型、文件大小、文件中的代码行数和非原创文件等信息。

6.3.2.2 特有内容分析

对于检材中出现的与样本中相同或相似的软件署名（包括开发者和所属单位）、数字水印、注释、废程序段和特异性错误等，应分析记录并在鉴定意见中体现。

6.3.2.3 资源文件分析

对于包含资源文件的检材和样本，应根据资源文件情况，采用以下适当方式进行相似性分析：

a) 对于图标、图片和音视频等外部资源文件，按照 GB/T 29361 的规定进行比对检验；
b) 对于包含图标、图片和音视频等资源文件的压缩文件，可释放资源文件后，按照 GB/T 29361 的规定进行比对检验。

6.3.2.4 结构分析

对于包含2层及以上目录结构的检材和样本，应分析其目录结构及文件分布的相似性。

6.3.2.5 代码分析

对于包含代码文件的检材和样本，应重点关注特有内容、代码执行逻辑与结构等内容，并采用以下适当方式进行相似性分析：

a) 对于空行、缩进和字符大小写等不影响程序执行的语法差异应予以排除或视为相同，排除或视为相同的规则应在鉴定记录中体现；
b) 对于注释内容宜分别进行排除前和排除后的相似性分析，排除规则应在鉴定记录中体现；
c) 对于文件名、变量名和函数名等不影响程序执行的名称差异可采用适当的规则进行替换，替换规则应在鉴定记录中体现；
d) 对于不影响程序执行的代码行或代码块的位置差异可采用适当的规则进行对齐，对齐规则应在鉴定记录中体现；
e) 对于以文本行方式比对的代码，可按对应顺序合并文件后进行代码行比对，合并顺序应在鉴定记录中体现。

6.3.2.6 数据库分析

对于包含数据库的检材和样本，可将数据库转储为sql脚本代码，分析数据库的库名、数据表名称、结构及数据、视图名称及定义、存储过程以及触发器等项目的相似性。

对于安装后预置了数据库数据的检材和样本，应重点比较预置数据的相似性。

6.3.2.7 其他内容分析

对于检材中与样本中其他相同或相似的内容，如执行流程图、测试用例、通讯端口情况和调试日志输出等，应分析并记录，必要时在鉴定意见中体现。

6.3.3 相似比例计算

6.3.3.1 结构相似比例的计算

6.3.3.1.1 以样本作为参考对象的结构相似比例计算公式见式（1）。

$$SS_k = \frac{F_c}{F_k} \times 100\% \cdots\cdots (1)$$

式中：

SS_{k} ——以样本作为参考对象的结构相似比例；

F_c ——包含文件夹结构比对后的同名文件数量；

F_k ——样本的文件数量。

6.3.3.1.2 以检材作为参考对象的结构相似比例计算公式见式（2）。

$$SS_q = \frac{F_c}{F_q} \times 100\% \cdots\cdots (2)$$

式中：

SS_q ——以检材作为参考对象的结构相似比例；

F_c ——包含文件夹结构比对后的同名文件数量；

F_q ——检材的文件数量。

6.3.3.2 文件相似比例

6.3.3.2.1 总体文件相似比例

6.3.3.2.1.1 以样本作为参考对象的总体文件相似比例计算公式见式（3）。

$$FS_k = \frac{D_c}{F_k} \times 100\% \cdots\cdots (3)$$

式中：

FS_k ——以样本作为参考对象的总体文件相似比例；

D_c ——完整性校验值相同的文件数量；

F_k ——样本的文件数量。

6.3.3.2.1.2 以检材作为参考对象的总体文件相似比例计算公式见式（4）。

$$FS_q = \frac{D_c}{F_q} \times 100\% \cdots\cdots (4)$$

式中：

FS_q ——以检材作为参考对象的总体文件相似比例；

D_c ——完整性校验值相同的文件数量；

F_q ——检材的文件数量。

6.3.3.2.2 排除非原创文件后的文件相似比例

6.3.3.2.2.1 以样本作为参考对象的排除非原创文件后的文件相似比例计算公式见式（5）。

$$\mathrm{FS}_{kn} = \frac{D_c - D_{nc}}{F_k - F_{\mathrm{kn}}} \times 100\% \cdots\cdots (5)$$

式中：

FS_{kn} ——以样本作为参考对象的排除非原创文件后的文件相似比例；

D_c ——完整性校验值相同的文件数量；

D_{nc} ——完整性校验值相同的非原创文件数量；

F_k ——样本的文件数量；

F_{kn} ——样本的非原创文件数量。

6.3.3.2.2.2 以检材作为参考对象的排除非原创文件后的文件相似比例计算公式见式（6）。

$$\mathrm{FS}_{qn} = \frac{D_c - D_{nc}}{F_q - F_{\mathrm{qn}}} \times 100\% \cdots\cdots (6)$$

式中：

FS_{qn} ——以检材作为参考对象的排除非原创文件后的文件相似比例；

D_c ——完整性校验值相同的文件数量；

D_{nc} ——完整性校验值相同的非原创文件数量；

F_q ——检材的文件数量；

F_{qn} ——检材的非原创文件数量。

6.3.3.3 代码相似比例

6.3.3.3.1 代码总体相似比例

6.3.3.3.1.1 以样本作为参考对象的代码总体相似比例计算公式见式（7）。

$$CS_k = \frac{C_c}{C_k} \times 100\% \cdots\cdots (7)$$

式中：

CS_k ——以样本作为参考对象的代码总体相似比例；

C_c ——代码相同行数；

C_k ——样本代码总行数。

6.3.3.3.1.2 以检材作为参考对象的代码总体相似比例计算公式见式（8）。

$$CS_q = \frac{C_c}{C_q} \times 100\% \cdots\cdots (8)$$

式中：

CS_q ——以检材作为参考对象的代码总体相似比例；

C_c ——代码相同行数；

C_q ——检材代码总行数。

注：对于空行、空格和字符大小写差异视为相同。

6.3.3.3.2 排除非原创内容后的代码相似比例

6.3.3.3.2.1 以样本作为参考对象的排除非原创内容后的代码相似比例计算公式见式（9）。

$$CS_{kn} = \frac{C_c - C_{nc}}{C_k - C_{kn}} \times 100\% \cdots\cdots (9)$$

式中：

CS_{kn} ——以样本作为参考对象的排除非原创文件后的代码相似比例；

C_c ——代码相同行数；

C_{nc} ——代码中非原创内容相同行数；

C_k ——样本代码总行数；

C_{kn} ——样本代码中非原创内容行数。

6.3.3.3.2.2 以检材作为参考对象的排除非原创内容后的代码相似比例计算公式见式（10）。

$$CS_{qn} = \frac{C_c - C_{nc}}{C_q - C_{qn}} \times 100\% \cdots\cdots (10)$$

式中：

CS_{qn} ——以检材作为参考对象的排除非原创文件后的代码相似比例；

C_c ——代码相同行数；

C_{nc} ——代码中非原创内容相同行数；

C_q ——检材代码总行数；

C_{qn} ——检材代码中非原创内容相同行数。

注：空行或仅有空白字符的行从行数中排除，仅存在多余空白字符、字符大小写差异，或其他不影响程序执行的名称或位置差异的行视为相同。

6.3.3.3.3 排除非原创内容及注释后的代码相似比例

6.3.3.3.3.1 以样本作为参考对象的排除非原创内容及注释后的代码相似比例计算公式见式（11）。

$$CS_{kna} = \frac{C_c - C_{nc} - C_a}{C_k - C_{kn} - C_a} \times 100\% \cdots\cdots (11)$$

式中：

CS_{kna} ——以样本作为参考对象的排除非原创内容及注释后的代码相似比例；

C_c ——代码相同行数；

C_{nc} ——代码中非原创内容相同行数；

C_a ——注释行；

C_k ——样本代码总行数；

C_{kn} ——样本代码中非原创内容行数。

6.3.3.3.3.2 以检材作为参考对象的排除非原创内容及注释后的代码相似比例计算公式见式（12）。

$$\mathrm{CS}_{q\mathrm{na}} = \frac{C_c - C_{\mathrm{nc}} - C_a}{C_q - C_{q\mathrm{n}} - C_a} * 100\% \cdots\cdots (12)$$

式中：

CS_{qna} ——以检材作为参考对象的排除非原创内容及注释后的代码相似比例；

C_c ——代码相同行数；

C_{nc} ——代码中非原创内容相同行数；

C_a ——注释行；

C_q ——检材代码总行数；

C_{qn} ——检材代码中非原创内容相同行数。

注：对于空行、空格、字符大小写、不影响程序执行的名称差异和位置差异等视为相同。

7 鉴定记录

7.1 基本要求

与检验有关的情况应及时、客观和全面地记录，保证检验过程和结果的可追溯性。

7.2 检材/样本信息

对于检材/样本，宜记录以下信息：

a) 软件的来源，如载体电子设备的类别、品牌型号、唯一性编号、性状或下载网址、远程访问的账号及口令等；
b) 软件载体或所处环境的照片；
c) 软件的名称、版本和大小等属性信息；
d) 软件的数字水印信息或签名信息；
e) 软件的完整性校验值；
f) 软件的运行环境；
g) 附属信息，如账号、密码等。

7.3 提取固定过程

对于检材/样本的提取固定过程，应记录以下信息：

a) 提取固定过程中所使用的仪器设备信息；
b) 远程提取的开始和结束时间（如适用）；
c) 提取固定结果的文件名和完整性校验值；
d) 固定过程录像文件的文件名和完整性校验值（如适用）。

7.4 检验过程

对于检材/样本的相似性检验过程，应记录以下信息：

a) 检材与样本在比较前的预处理过程；
b) 检材与样本的相同及相似部分；
c) 6.3.2.2 中检出的特有内容；
d) 6.3.2.5 中的处理规则；
e) 6.3.2.7 中检出的其他内容；
f) 相似比例的计算过程。

8 鉴定意见

8.1 鉴定意见分类

软件相似性鉴定意见应分为以下四种：

a) 软件相同；
b) 软件相似；
c) 软件不相似；
d) 无法判断。

8.2 鉴定意见判断依据及表述

8.2.1 软件相同

判断依据：检材与样本（安装文件、代码文件和各组成文件等）使用GB/T 29361比较结果相同。
鉴定意见表述为：检材与样本（列出对应的文件）相同。

8.2.2 软件相似

判断依据：检材与样本中存在部分文件的完整性校验值相同或文件的内容存在相同部分。
鉴定意见表述为：检材与样本相似，并列出相似比例。
附加要求如下：
a) 应分别列出以检材为参考对象和以样本为参考对象时，检材与样本的相似比例；
b) 如检材与样本包含多种文件类型，可分别列出各类型文件的相似比例；
c) 对检材与样本中存在相同或相似的部分应进行说明；
d) 不宜出现“实质性相似”表述。

8.2.3 软件不相似

判断依据：检材与样本中不存在完整性校验值相同的文件且文件内容不存在相同部分。
鉴定意见表述为：检材与样本不相似。

8.2.4 无法判断

判断依据：检材与样本不具备检验条件，或在进行了充分的检验后仍无法判断是否相似。
鉴定意见表述为：无法判断检材与样本是否相似。

参 考 文 献

[1] GB/T 29360—2023 法庭科学 电子数据恢复检验规程
[2] GB/T 36303—2018 印刷数字水印
[3] GA/T 756—2021 法庭科学 电子数据收集提取技术规范
[4] GA/T 976—2012 电子数据法庭科学鉴定通用方法
[5] GA/T 1175—2014 软件相似性检验技术方法
[6] SF/Z JD0400001—2014 电子数据司法鉴定通用实施规范